==========
Language in India www.languageinindia.com **ISSN 1930-2940 Vol. 20:1 January 2020**
==========

MACHINE TRANSLATION – YESTERDAY, TODAY AND TOMORROW

(இயந்திர மொழிபெயர்ப்பு – நேற்று, இன்று, நாளை)

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

==========

ABSTRACT

"Machin Translation - Yesterday, Today and Tomorrow" is a monograph written in Tamil to introduce the deferent dimensions of machine translation to the Tamil readers. It has nine chapters including introduction and conclusion.

Chapter 1: Introduction

The introduction describes the aim and scope of machine translation. It discusses about the general characteristics of machine translation, lists the need for machine translation and introduces different types of machine translation, lists different types of grammatical formaiims used for machine translation as well as preliminary steps to machine translation.

Chapter 2: Mile stones in Machine Translation

This chapter discusses about the history of machine translation explained by Hutchins, Somers and others. Hutchins and Somers describes about the history of machine translation by diving the stages of MT development into four or five stages. The first stage covers the period from 1933 to 1956. The second stage covers the period from 1956 to 1966. Next comes the report of ALPAC which criticized the MT attempts and the futility of the efforts. The fourth stage covers the period from 1967 to 1976. This is followed by a description on translation tools and translators' workstation. "Research since 1989" is described afterwards. Next mile stone is "operational and commercial system since 1990". "MT on Internet" becomes the topic of next stage.

Chapter 3: Some Important Machine Translation Systems

This chapter discusses about the different types of MT systems built in world arena. The following MT systems have been discussed briefly: GAT:Georgetown automatic translation, CETA automatic translation center research, TAUM research in University of Montreal, ALP automatic

language processing, SYSTRAN, LOGOS, SUSY, METEO, Weinder Communication Corporation, SPANAM, CULT: Chinese University Translator, ALPS automated language processing systems., MT development in Japan, Ariane (GETA), Eurotra, METAL, Rosetta and DLT.

Chapter 4: MT Research in India

This chapter discusses about the major attempts to build Machine transplantation systems in India. The MT systems such as Anusaraka, Anglabharati, Anubharati, Anglabharati II, Mantra MT system, Anuvadak MT system, MT system based on ULN, Matra MT system, English-Kannada MT system and Indian language to Indian Language machine translation systems have been described. This is followed by a description of Tamil oriented MT systems. Tamil-Rusian MT system, UNL based MT system for Tamil, Machine aide to translate Linguistics text books in English into Tamil and English-Tamil machine Translation system using statistical methods.

Chapter 5: Corpus Creation

This chapter talks about the corpora creation. The different corpora created for English language is discussed initially. This is followed by a description on Indian language corpora creation.

Chapter 6: POS Tagging

This chapter talks about the POS tagging of texts. The importance of POS tagging, creation of TAG sets, different types of TAG sets available at the international arena as well as in Indian language arena have been discussed.

Chapter 7: Syntactic Parsing

This chapter discusses about syntactic parsing. The different types of syntactic parsing and the attempts to parse Indian languages including Tamil have been dealt in this chapter.

Chapter 8: Conclusion

இயந்திர மொழிபெயர்ப்பு – நேற்று இன்று நாளை

MACHINE TRANSLATION – YESTERDAY, TODAY AND TOMORROW

முனைவர் ச. இராசேந்திரன்

பணிநிறைவுற்ற பேராசிரியர், தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்

தற்போது

வருகைதரு பேராசிரியர் அமிர்தா விஷ்வ வித்யபீடம், கோயம்பத்தூர்

rajushush@gmail.com

கோயம்புத்தூர்

January 2020

எனது உரை

இயந்திர மொழிபெயர்ப்பு ஆய்வில் எனது பயணம் நீண்ட பயணம். இந்திய இயந்திர மொழிபெயர்ப்பு முயற்சிகளின் ஆரம்ப கட்டமான அனுசரகா உருவாக்கத்திலிருந்து (1989) இன்றைய புள்ளியல் அடிப்படை இயந்திர மொழிபெயர்ப்புவரை நான் பயணித்து வருகின்றேன். தஞ்சாவூர் தமிழ்ப் பல்கலைக்கழத்தில் இது குறித்த சில ஆய்வுத்திட்டங்களை நான் செய்துவரும்போது இயந்திர மொழிபெயர்ப்பு பற்றி படிப்பதற்கும் அதன் நுணுக்கங்களை அறிந்துகொள்வதற்கும் ஏராளமான வாய்புகள் கிடைத்தன. அக்காலகட்டத்தில் குறிப்பாக ஹட்சின்ஸும் சோமர்ஸும் (Hutchins and Somers, 1992) எழுதிய இயந்திர மொழிபெயர்ப்பு அறிமுகம் (Introduction to Machine Translation) என்ற நூல் நல்ல நோக்கீட்டு நூலாக அமைந்தது. இயந்திர மொழிபெயர்ப்பின் வரலாறும் அன்றைய காலகட்ட இயந்திர மொழிபெயர்ப்பு முயற்சிகளும் பற்றி அதன் மூலம் அறிய முடிந்தது. சிஸ்ட்ரான், யுரொட்ரான், கெதா, லூசி போன்ற ஐரோப்பிய அமேரிக்க ஜப்பானிய மொழிபெயர்ப்பு ஒழுங்குமுறைகள் எவ்வாறு இயந்திர மொழிபெயர்ப்பு செயற்பாங்கைத் தொகுதிவாரியாகப் பிரித்துக்கொண்டு மொழியியல் கொள்கைகளையும் இலக்கண விதிகளையும் ஆழ்ந்து ஆய்ந்து மொழிபெயர்ப்புச் சிக்கல்களை அணுகுகி அவற்றைத் தீர்க்கின்றார்கள் என்பதுபோன்ற தகவல்களை அறிந்து கொண்டு இயந்திர மொழிபெயர்ப்பில் ஈடுபடுவது நல்லது. இந்திய இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மொழியியலின் முக்கியத்துவத்தையும் பயன்பாட்டையும் கொள்ளைகளையும் அறியாமலோ உணராமலோ அல்லது தேவையில்லை என்று ஒதுக்கிகொண்டோ மொழியியலை 'கறிவேபிலை' போன்று கையாளுகின்றனவோ என்று நான் எண்ணுவது உண்டு.

தற்போது மொழிபெயர்ப்பு அணுகுமுறை முற்றிலும் மாறிவிட்டது விதி அடிப்படையிலான அணுகுமுறையும் விதி அடிப்படையிலான இயந்திரம் கற்றல் அணுகுமுறையும் முற்றிலுமாகக் கைவிடப்பட்டுள்ளது. சாம்ஸ்கியின் கருத்தான தரவு அடிப்படையிலான இலக்கணம் முழுமையான இலக்கணமாக அமையாது என்பது முற்றிலும் மறக்கப்பட்டு இணையான தரவுகளைக் கொண்டு ஆழக்கற்றல் முறையில் நரம்புப் பின்னல் அணுகுமுறை அடிப்படையில்

மொழிபெயக்கும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. இதன் பலனாகக் கணி அறிவியலார் மொழி பற்றிய அறிவோ இலக்கண அறிவோ இன்றி இருமொழிகளின் இணைத் தரவுத்தொகுதியை வைத்துக்கொண்டு புள்ளியியல் அடிப்படையில் மொழிபெயர்க்க முயற்சிக்கின்றனர். அவர்கள் இவ்வணுகுமுறையில் மொழியியல், இலக்கணம், அகராதி இவற்றைச் சார்ந்த கடுமையான வியர்வை சிந்தும் செல்பாடுகளைச் செய்யத் தேவையில்லை. அவர்கள் பாரம் முற்றிலுமாக குறைந்து கண்ணை மூடிக்கொண்டு இதற்கான மென்பொருளைப் பயன்படுத்தி மொழியியலார் தயவை நாடாமல் தன்னிச்சையாக இயந்திர மொழிபெயர்ப்பு செய்கின்றனர். இது உண்மையிலேயே இயந்திரம் அடிப்படையிலான மொழிபெயர்ப்பு தான். எந்த மொழியியல் அறிவும் இலக்கண அறிவும் அகராதி அறிவும் இவ்வணுகு முறைக்குத் தேவையில்லை. சுருக்கமாகக் கூறப்போனால் இது ஒரு கண்மூடித்தனமான அணுகுமுறை. இதில் ஏற்படும் தவறுகளை நாம் கணிக்கவோ தவறைத் திருத்துவதற்கான வழிமுறைகளை மேற்கொள்ளவோ இயலாது. என்ன நடக்கிறது என்பது ஆண்டவனுக்குத்தான் வெளிச்சம் என்பது போல மொழிபெயர்ப்பு கணிப்பொறிக்குத்தான் வெளிச்சம். இந்தக்காலமும் கடந்துப்போகும் என்று காத்திருப்போம். உமிக்கரியிலிருந்து பற்பசைக்கு மாறிய நாம் மீண்டும் உமிக்கரிக்கே திரும்புவது போன்று ஒரு சுற்று கடந்து மீண்டும் விதி அடிப்படையிலான அணுகுமுறைக்கே திரும்பும் நாள் அதிகம் இல்லை.

இந்நூல் இயந்திர மொழிபெயர்ப்பில் 1998-இலிருந்து இன்றையவரையிலான எனது பயணத்தின் ஒரு தொடர்கதைதான். நான் படித்தவற்றை எல்லாம் எனது தமிழ் மொழியியல் மாணவர்களை மனதில் இருத்தி தமிழில் குறிப்புகள் எடுத்து எனது மடிக்கணினியில் உள்ளீடு செய்து வந்தேன். இதுகுறித்து மாணவர்களுக்கு வேண்டிய இயந்திர மொழிபெயர்ப்பு குறித்த பாடக்குறிப்புகளையும் கணிபொறியில் உள்ளீடு செய்து வந்தேன். இது குறித்து பல கட்டுரைகளையும் எழுதி கணினியில் உள்ளீடு செய்து வந்தேன். இவற்றையெல்லாம் ஒன்று திரட்டி தமிழுக்கு இயந்திர மொழிபெயர்ப்பு குறித்து ஒரு மிகப்பெரிய நூலைப் படைக்கவேண்டும் என்ற எனது நீண்டநாள் அவாவை வேறு வழியின்றி அறைகுறையாக இங்கு நிறைவு செய்கின்றேன். இதுவரை நான் இயந்திரமொழிபெயர்ப்புகுறித்து எழுதியவைகளைத் திருத்தி வெளியிடுவதற்கு எனக்கு காலமும் இல்லை நிதியும் இல்லை. இயந்திர மொழிபெயர்ப்பை

உள்ளடக்கிய இயற்கைமொழியியல் ஆய்வில் மொழியியலின் இன்றியமையாமை ஏற்றுக்கொள்ளப்பட்ட ஒன்று. தமிழ் நாட்டில் மொழியியலின் முக்கியதுவம் சரியாக கவனக்குவிப்பு செய்யப்படவில்லை என்ற ஆதங்கம் எனக்கு எப்பொழுதும் உண்டு. மொழியியல் ஒரு தேவையில்லாத பாடமாகத்தான் தமிழ் நாட்டில் இருந்துவருகின்றது. எனவே தற்போதைய எனது வீண் முயற்சி என்று எனக்குத் தெரியும். இதுதான் தற்போதைய இந்நூல் வரைவின் வரலாறு. நூல் மேலும் செப்பனிடப்படவேண்டும். இதைக் காகிதப் பதிப்பாக வெளியிட யாராவது முன்வந்தால் அம்முயற்சியை மேற்கொள்ளலாம். இதுவரை எனது மடி(க்கணி)யில் தூங்கிக்கொண்டிருந்த இந்த நூல்வரைவு இன்றுமுதல் இணையத்தில் தூங்கட்டும்.

இந்நூலை Langaue in India எனற மின் திங்களிதழிலில் வெளியிட ஒப்புக்கொண்ட பேராசிரியர் திருமலை அவர்களுக்கு எனது நன்றியைத் தெரிவித்துக் கொள்கின்றேன்.

ச.இராசேந்திரன்

பொருளடக்கம்

இயல் 1

முன்னுரை

1.1. அறிமுகம்

பாரதத்தில் பன்மொழி பேசும் மக்கள் உள்ளனர். மொழி, இனம், பண்பாடு மற்றும் கலாச்சாரத்தால் பல்வேறு மக்களின் எண்ணங்கள், கருத்துக்கள் பிறமொழி பேசும் மக்களிடையே சென்றடைவது அவசியமாகும். அதுபோல் ஒருமொழி பேசும் மக்களின் பண்பாடு, கலாச்சாரம், இலக்கியம் போன்றவைகளும் பிறமொழி பேசும் மக்களிடையே சென்றடைய வேண்டும். அப்பொழுதுதான் மக்களிடையே ஒற்றுமை உணர்ச்சியும் தேசிய ஒருமைப்பாடும் நிகழும். ஒரு மொழியில் உள்ள செய்திகள் பிற மொழிகளில் மாற்றம் செய்யப்படுவது மிகவும் தேவையாகும். இதுபோல் உலக நாடுகளில் ஒரு நாடு பிற நாடுகளிலிருந்து தனித்து இருக்க இயலாது. தொழிற்நுட்ப மேம்பாடு, பொருளாதார மேம்பாடு, சமூக மேம்பாடு போன்றவைகளுக்கு ஒன்றோடு ஒன்று தொடர்பு கொண்டு கூட்டு முயற்சியில் மனித வாழ்வியலில் மேம்பாட்டினைக் கருத்தில் கொள்ள வேண்டும். ஒரு நாட்டின் கலாச்சாரம் பண்பாடு போன்றன பிற நாடுகளுக்கும் பரவச்செய்யப்பட வேண்டும். ஒரு நாட்டின் இலக்கியச் செல்வங்கள் பிற நாடுகளைச் சென்றடைய வேண்டும். அப்பொழுதுதான் ஒரு நாட்டின் தொன்மை, தனித்துவம், வளம் போன்றவைகள் பிற நாடுகளுக்குத் தெரியவரும். எனவே ஒரு நாட்டில் உருவாக்கப்படும் படைப்புகள் பிற நாடுகளுக்கு எளிதில் சென்றடைய வேண்டுமெனில் மொழிபெயர்ப்புப் பணி அவசியமாகும். நாம் தகவல் யுகத்தில் வாழ்ந்து வருகின்றோம். செய்திகள் மலையெனக் குவிகின்றன. இச்செய்திகளை உடனுக்குடன் மொழிபெயர்க்க வேண்டியது அவசியமாகிறது. வாழ்வில் எல்லா நிலைகளிலும் பயன் தரக் கூடிய கணிப்பொறி மொழி பெயர்ப்புக்குப் பயன்படுத்த மேற்கொள்ளப் பட்டுள்ளன.

மனித மொழிபெயர்ப்பைக் காட்டிலும் இயந்திர மொழிபெயர்ப்பு விரைவானதாகும். மனித மொழிபெயர்ப்பில் கால விரயமும் களைப்பும் ஏற்படும். சரியாக நடைமுறைப்படுத்தப்பட்டால் இயந்திர மொழிபெயர்ப்பு மனித மொழிபெயர்ப்பிற்கு இணையாகச் சிறப்பாக அமையும். இயந்திர மொழிபெயர்ப்பின் போது பல துணைக்கருவிகள் உருவாக்கப்படுகின்றன. இவை மொழியாய்விற்கும் பலவித மொழிப் பயன்பாடுகளுக்கும் உதவி புரியும். கணிப்பொறி வழி மொழிபெயர்ப்பதன் தேவையையும் அதன் வகைகளையும் அவற்றிற்கான முயற்சிகளையும் எல்லைகளையும் ஆய்வது தான் இப்பகுதியின் நோக்கம் ஆகும்.

அன்றிலிருந்து இன்று வரை மொழிபெயர்ப்பு இன்றியமையாத மொழிச் செயல்பாடாக அமைகின்றது. பல மொழிகள் பேசப்படுகின்ற இந்தியா போன்ற நாடுகளில் மொழிபெயர்ப்பு ஒரு

பொதுவான நிகழ்வாக நடந்துகொண்டிருக்கின்றது. ஒரு மொழியில் எழுதப்படுகின்ற மற்றும் பேசப்படுகின்ற எந்தத் தகவலும் செய்தியும் உடனுக்குடன் வேற்று மொழிகளில் மொழிபெயர்க்கப்படவேண்டிய கட்டாயம் நம்மிடையே இருந்து கொண்டிருக்கின்றது. பன்மொழிச் சூழலைப் பிரதிபலிக்கும் இந்தியா போன்ற நாடுகளில் மொழி பெயர்ப்பு ஒரு அன்றாடப் பணியாக நடந்து கொண்டிருக்கின்றது. மொழிபெயர்ப்பாளர்கள் எல்லாக் காலகட்டத்திலும் முக்கித்துவம் வாய்ந்தவர்களாக இருந்து வந்திருக்கின்றார்கள்; மொழிபெயர்ப்பும் எல்லாக் காலகட்டத்திலும் முக்கியத்துவம் வாய்ந்த செயல்பாடாக இருந்து வருகின்றது. நமது சிந்தனைகள் ஒலிக்குறியீடுகளாகவோ எழுத்துக் குறியீடுகளாகவோ மாற்றப்படுவதும் ஒரு வித பெயர்ப்பே. கணிப்பொறியின் வரவால் ஒரு மொழி உரைகளை மற்றொரு மொழி உரைகளாக மாற்றும் பணி மற்றொரு பரிமாணத்தை நோக்கிப் பயணிக்கின்றது. மனிதச் செயல்பாடான மொழிபெயர்ப்பைக் கணிப்பொறிச் செயல்பாடக மாற்றும் கட்டாயத்திற்கு நாம் தள்ளப்பட்டுள்ளோம். இதன் காரணமாக இயந்திர மொழிபெயர்ப்பு இன்றைய காலகட்டத்தில் இன்றியமையாத செயல்பாடாகக் கருதப்பட்டு அதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

மொழிபெயர்ப்புச் செயல்பாட்டின் போது ஒரு மொழிபெயர்ப்பாளர் அறிந்திருக்கவேண்டிய தகவல்கள் யாவும் கணிபொறிக்கு தந்தால்தான் இது சாத்தியமாகும். மொழிபெயர்ப்பவர் மூலமொழி அமைப்பு பற்றியும் இலக்குமொழி இலக்கண அமைப்பைப் பற்றியும் அறிந்திருக்கவேண்டும். எவ்வாறு மூலமொழியின் இலக்கண அமைப்பும் இலக்குமொழியின் இலக்கண அமைப்பும் வேறு படுகின்றன என அறிந்திருக்கவேண்டும். மூல மொழிச் சொற்களுக்கு இணையான இலக்கு மொழிச் சொற்கள் என்ன? மூல மொழித் தொடர்களுக்கு இணையான இலக்கு மொழித் தொடர்கள் என்ன? மூல மொழி வாக்கியங்களுக்கு இணையான இலக்குமொழி வாக்கியங்கள் என்ன என்பன மொழிபெயர்ப்பாளர் அறிந்திருக்கவேண்டும். மூலமொழியை இலக்கு மொழியாக மாற்றப் பணித்துள்ள கணிப்பொறியும் இவ்வறிவைப் பெற்றிருக்கவேண்டும். மூல மொழியின் மற்றும் இலக்குமொழியின் உருபனியல் அமைப்பு, சொல்லியல் அமைப்பு, தொடரியல் அமைப்பு, பொருண்மையியல் அமைப்பு என்பன பற்றியும் எவ்வாறு இவ்வமைப்புகள் மூல மொழியிலும் இலக்கு மொழியிலும் வேறு படுகின்றன என்பது பற்றியும் கணிப்பொறிக்கு அறிவுறுத்த

வேண்டும். மொழி உரைகள் வெறும் எழுத்துக்களின் கோர்வையல்ல, எழுத்துக்கள் உருபன்களாகவும் உருபன்கள் சொற்களாகவும் சொற்கள் தொடர்களாகவும் தொடர்கள் வாக்கியங்களாகவும் பகுப்பாய்வு செய்யப்பட்டுதான் எழுதுக் கோர்வைகள் பொருள் பெறுகின்றன என்ற நமது அறிவைக் கணிப்பொறிக்குத் தரவேண்டும். எனவே தமிழ்த்தரவுத் தொகுதிகளைச் சொல்வகை அடையாளப்படுத்தல் மற்றும் பெயர்த்தொடர் உறுப்பு பகுத்தல் என்பன இயந்திர மொழி பெயர்ப்பின் முக்கிய செயல்பாடுகளாக மாறுகின்றது.

மூலமொழி அமைப்பு பற்றிய அறிவையும் இலக்கு மொழி அமைப்பு பற்றிய அறிவையும் அவைகள் எவ்வாறு வேறு படுகின்றன என்ற அறிவையும் அதன் அடிப்படையில் மூல மொழி உரைகளை இலக்கு மொழி உரைகளாக மாற்றும் விதிகளையும் கணிப்பொறிக்குத் தந்து கணிப்பொறியை மொழிமாற்றம் செய்யப் பணித்தால் கணிப்பொறி அப்பணியைச் செவ்வனே செய்யும் என்பது எதிர்பார்ப்பாகும். ஆனால் எதிர்ப்பார்ப்புக்கும் சாத்தியத்திற்கும் விளைவிற்கும் இடையில் பெரிய இடைவெளிகள் உள்ளன. கணிப்பொறிக்கு விதிகளைச் சொல்லித் தருவதைவிட விதிகளைத் தானாகவே உணரச் செய்யும் செயல்பாடு எளிதானதும் சாத்தியமானதாக அமைகின்றது. இத்தகைய சூழலில் புள்ளியியல் செயல்பாடு முக்கியத்துவம் பெறுகின்றது. கணிப்பொறிக்கு எவ்வாறு செய்யவேண்டும் என்ற பயிற்சியை அளித்துவிட்டால் அது தானே நாம் விரும்பும் செயல்பாட்டைத் தவறின்றிச் செய்யத் தொடங்கும். தேவையான அளவு சொல் எண்ணிக்கையுள்ள தமிழ்த் தரவுத் தொகுதிகளைச் சேகரித்து சொல் வகை அடையாளப் படுத்தலையும் (Parts of speech (POS) tagging) தொடர்ப் பிரித்து அடையாளப்படுத்தலையும் (Chunking) செய்து கணிப்பொறிக்குத் தந்தால் அது அச்செயற்பாடுகளைத் தானாகவே கற்றுக்கொண்டு புதிய தரவுத்தொகுதிகளைச் சொல்வகைக்கு அடையாளப்படுத்துவதையும் தொடர்களாகப் பிரிப்பதையும் (chunking) தானாகவே செய்யும். இதன் அடிப்படையில் தான் தரவுத் தொகுதிகளைச் சேகரித்து சொல்வகை அடையாளப்படுத்துவதையும் தொடர்பிரித்து அடையாளப்படுத்துவதையும் மனித முயற்சியால் செய்து அதைக் கணிப்பொறிக்குக் கற்பித்து (machine learning) அந்த அறிவைப் பயன்படுத்தி கணிப்பொறி புதிய தரவுத் தொகுதிகளைச் சொல்வகைக்கு அடையாளப்படுத்தவும் தொடர்பிரித்து அடையாளப்படுத்தவும் செய்யும்படிக்கு செயற்படுத்த இயலும். இச்செயற்பாடுகள் இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு மொழிபெயர்ப்பு ஒழுங்குமுறை என்ற மைய அரசின்

கருத்துப்பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் நிதி நல்கையில் நடைபெற்றுவரும் திட்டத்தின் கீழ் உருவாக்கப்பட்ட இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் ஒரு பகுதியாக அமைகின்றது.

தமிழ் மொழியின் தொழில்நுட்ப முன்னேற்றத்திற்கு வேண்டி அண்ணா பல்கலைக்கழகத்தில் உருவாக்கப்பட்ட RCILT (Resource Centre for Indian Languages-Tamil) இத்தகைய முயற்சியில் ஈடுபட்டு தமிழுக்கு உருபனியல் பகுப்பாய்வி, சொல்வகைப்பாட்டு அடையாளப்படுத்தி என்பனவற்றை ஓரளவுக்கு வெற்றிகரமாக உருவாக்கியுள்ளது. இருப்பினும், இதைக் கடந்து அடுத்த நிலைக்குச் செல்லும் தொடரியல் பகுப்பான் உருவாக்கும் முயற்சி தொடக்க நிலையிலேயே உள்ளது. தமிழில் பொருள்கோளின் பெரும்பகுதி உருபனியல் நிலையிலேயே செய்யப்படுகின்றது. ஏனென்றால் தமிழ் ஒரு வினையின் பங்கெடுப்பாளரின் செயற்பாங்குகளை வெளிப்படுத்த வேற்றுமை உருபுகளையே கூடுதலாகப் பயன்படுத்துகின்றது. இதன் காரணமாகத் தமிழ் ஒத்தறி அடிப்படையில் ஒரு சொல் சுதந்திரமான மொழியாக (Free Word Order Languge) உள்ளது. தமிழில் எழுவாய்த் தொடர், செயப்படுபொருள் தொடர் போன்ற பெயர்த் தொடர்களை எவ்வித பொருண்மை மாற்றமும் இன்றி வரிசை மாற்றி அமைக்க இயலும். இதன் காரணமாக ஆங்கில மொழியில் உள்ளது போன்ற கலவைத் தன்மையான தொடர் அமைப்பு தமிழுக்கு இல்லை. இருப்பினும், ஒரு தொடருக்குள் வரும் சொற்களும் சொலுக்குள் வரும் உருபன்களும் வரையறுக்கப்பட்ட நிரல்களிலேயே வருகின்றன. எனவே தொடர்களின் எல்லைகளையும் (boundary of phrases) எச்சத் தொடர்களின் எல்லைகளையும் (boundary of clauses) அறிந்து கொள்ள வேண்டியது இயந்திர மொழிபெயர்ப்பு போன்ற செயல்பாடுகளுக்கு மிக அவசியம் ஆகும். ஏனென்றால் இயந்திர மொழிபெயர்ப்பின் போது மூல மொழியின் அமைப்பு இலக்கு மொழியின் அமைப்பாக மாற்றப்பட வேண்டும். இம்மாற்றத்திற்கு ஆழமில்லாப் பகுப்பாய்வு (shallow parsing) அவசியமாகும். எனவே தொடரியல் பகுப்பான் தமிழுக்குத் தேவையே இல்லை என்று ஒதுக்கி விட முடியாது. தமிழில் ஒரு வாக்கியதிற்குள் வரும் உறுப்புகளை வரிசை மாற்றம் செய்ய இயலும் என்றாலும் ஒரு தொடருக்குள் வரும் உறுப்புகள் ஒரு குறிப்பிட்ட நிரலிலேயே வருகின்றன. மேலும் ஒரு தொடரில் உள்ள உறுப்புகளை மற்ற தொடருக்குள் நகர்த்த இயலாது; தொடர்களை அல்லது தொடர்களின் உறுப்புகளை நகர்த்துவதில் தமிழும் பல கட்டுப்பாடுகளை வெளிப்படுத்துகின்றது. இதன் காரணமாகப் பொருள்கோளுக்கு (semantic interpretation) தொடரியல் பகுப்பான் மிக முக்கியதுவம் வகிக்கின்றது. உருபனியலைக்

கடந்த பல பொருண்மை மயக்கங்களைத் தொடரியல் பகுத்துக்குறிப்பான்தான் தீர்த்து வைக்க இயலும்.

மேற்கூறிய இயந்திர மொழிபெயர்ப்புச் சிக்கலையும் இயந்திர மொழிபெயர்ப்பின் எல்லைகளையும் கருத்தில் கொண்டு மொழிபெயர்ப்பிற்காக பன்மாதிரியான முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன.

1.2. கணிப்பொறி வழி மொழி பெயர்ப்பின் தேவை

ஒரு மொழியில் உள்ள தகவலை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதையே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழி பெயர்ப்பில் தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியின் இயல்புத் தன்மை மாறாது பாதுகாத்தல் வேண்டும். தகவல் எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Language) என்கிறோம். தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகின்றதோ அந்த மொழியை இலக்குமொழி (Target Language) என்கிறோம். இத்தகைய மொழி பெயர்ப்புகளை மனித முயற்சியால் மட்டும் மேற்கொள்வது என்பது கடினமானப் பணியாகும். எனவே மொழிபெயர்ப்புப் பணிகளைக் கணிப்பொறி வழி மேற்கொள்வதற்கான கட்டாயம் ஏற்படுள்ளது. கணிப்பொறி வழி மொழிபெயர்ப்பின் தேவைகளும் கட்டாயங்களும் கீழேப் பட்டியலிடப்பட்டுள்ளன.

- உலகளவில் தகவல்கள் அன்றாடம் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழிபெயர்த்தல் அவசியமாகிறது.
- கணிப்பொறி வழி மொழிபெயர்ப்பு மனித உழைப்பைக் குறைக்கிறது.
- குறைந்த நேரத்தில் நிறைய மொழி பெயர்ப்புப் பணிகளைச் செய்ய இயலுகிறது. இதனால் காலவிரயம் தவிர்க்கப்படுகிறது.
- தற்போதைய மொழி பெயர்ப்புப் பணிகளுக்கு ஏற்ற வகையில் மொழிப் பெயர்ப்பாளர்கள் இல்லை.
- மனித முயற்சியில் மொழிபெயர்கின்ற பொழுது மொழி பெயர்ப்பாளர்கள் மூலமொழியிலும் இலக்கு மொழியிலும் புலமை உள்ளவர்களிடம் கருத்துக்களைப் பெற்று அதை வழியமைப்பு மொழியில் எழுதி மொழிபெயர்ப்பு செய்யலாம்.

1.3. இயந்திர மொழிபெயர்ப்பின் வகைகளும் அணுகுமுறைகளும்

இயந்திர மொழிபெயர்ப்பில் மனித முயற்சியின் பங்கு என்ன என்பதன் அடிப்படையில் கணிப்பொறி வழி மொழிபெயர்ப்பினை பின்வருமாறு பகுக்கலாம்:

1. கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு

(CAMT – Computer Assisted Human Translation)

2. மனித உதவியின் கணிப்பொறி மொழிபெயர்ப்பு

(HAMT-Human Assisted Machine Translation)

3. முற்றிலும் தானியங்கி மொழிபெயர்ப்பு

(Automatic Machine Translation)

1.3.1.கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனிதமொழி பெயர்ப்பில் மனித மொழிப் பெயர்ப்பாளர்களுக்கு முக்கியத்துவம் தரப்படுகின்றது. அவர்கள் மொழிப்பெயர்ப்புப் பணியில் ஈடுபடும் பொழுது ஒரு மொழியில் உள்ள சொற்களுக்கு நிகரான பிறமொழிச் சொற்களைத் தெரிந்துகொள்ளவும் அச்சொற்களின் பொருள் பயன்பாடு ஆகியவற்றை எளிமையாகத் தெரிந்துகொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு, தொடரமைப்பு போன்றவைகளுக்கு நிகரான பிறமொழி அமைப்புகளை விரைவாகத் தயார் நிலையில் பெறுவதற்கும் கணிப்பொறி உதவுகிறது. இவ்வகை மொழிபெயர்ப்பில் மொழிப்பெயர்பாளர்கள் தான் முதல் நிலையில் செயலாற்றுவர்; அவர்களுக்கு உதவும் நிலையில் கணிப்பொறி செயலாற்றும்.

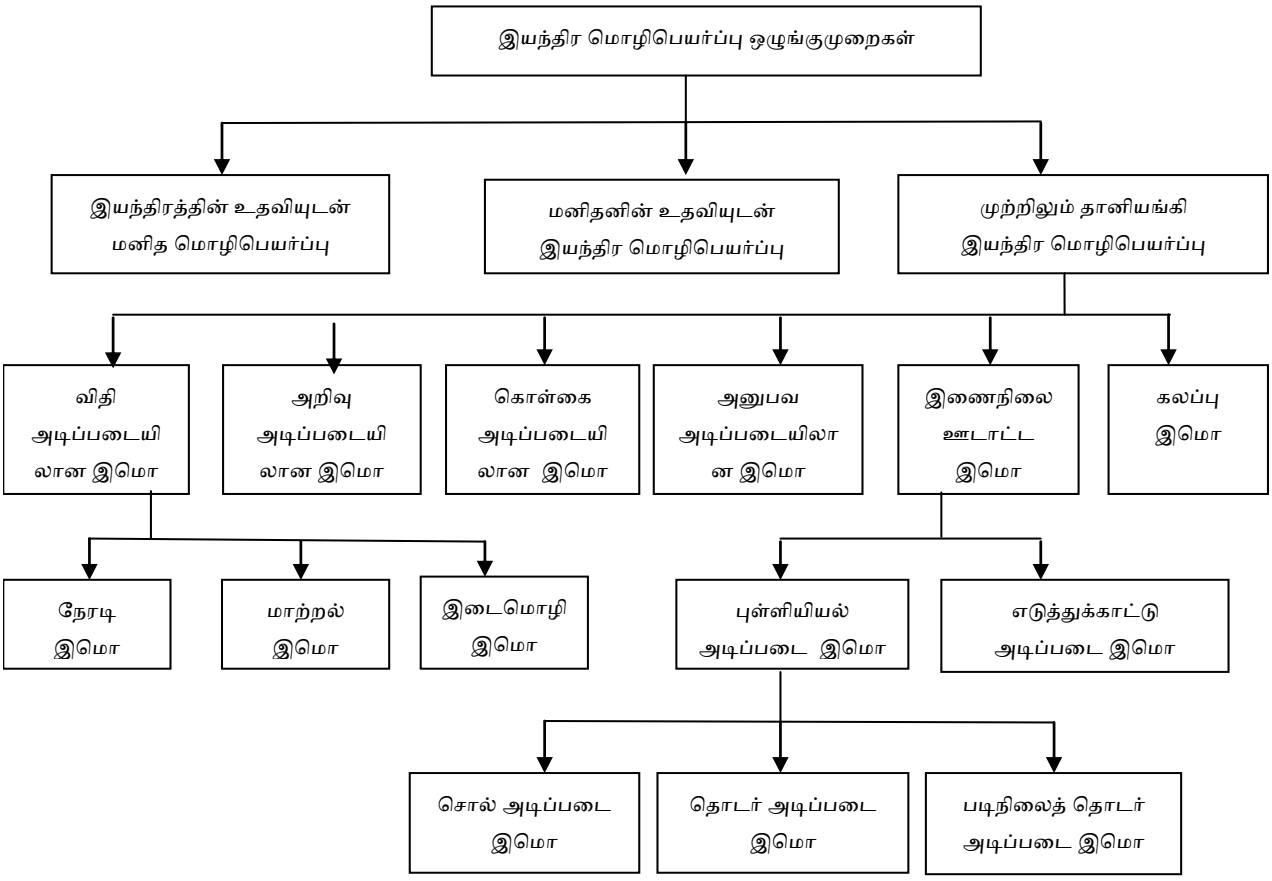
1.3.2.மனித உதவியுடன் கணிப்பொறி மொழிப்பெயர்ப்பு

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிப்பெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிப்பெயர்ப்புப் பணிகளைச் செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ள நகலுக்கு மேலாக ஏதாவது செய்திகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட மூவகை மொழிப்பெயர்ப்பு வகைகளில் மனித உதவியுடன் செயல்படும் இயந்திர மொழி பெயர்ப்பு தான் நடைமுறையில் பரவலாக உள்ளது.

1.3.4.முற்றிலும் தானியங்கி மொழிபெயர்ப்பு

முற்றிலும் இயந்திர மொழிபெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல இதற்கு எண்ணற்ற மொழியில் செயல்பாடுகளும் மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கப்பட்ட வேண்டி உள்ளது. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்பில் முற்றிலும் தானியங்கப்படுத்தும் முயற்சி இலக்காகக் கொள்ளப்பட்டுள்ளது.

இது தவிர இயந்திர மொழிபெயர்ப்பில் பயன்படுத்தப்படும் வழிமுறைகளின் அடிப்படையிலும் இயந்திர மொழிபெயர்ப்பைப் பலவிதமாகப் பகுக்கலாம். பொதுவாக, இயந்திர மொழிபெயர்ப்பு ஏழு பரந்த வகைகளாக வகைப்படுத்தப்படுகிறது: விதி அடிப்படையிலான, புள்ளிவிவர அடிப்படையிலான, கலப்பின அடிப்படையிலான, எடுத்துக்காட்டு அடிப்படையிலான, அறிவு சார்ந்த, கொள்கை அடிப்படையிலான மற்றும் ஆன்லைன் ஊடாடும் அடிப்படையிலானது அணுகுமுறைகள். முதல் மூன்று இயந்திர மொழிபெயர்ப்பு அணுகுமுறைகள் மிகவும் பரவலாகப் பயன்படுத்தப்படும் மற்றும் முந்தைய முறைகள். இந்த அணுகுமுறைகள் அனைத்தையும் பயன்படுத்தி பலனளிக்கும் முயற்சிகள் இருந்தன என்பதை இலக்கியம் காட்டுகிறது. ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கும் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்கும் முயற்சிகள் மேற்கொள்ளப்பட்டன; இன்றும் மேற்கொள்ளப்பட்டு வருகின்றன. தற்போது இயந்திரமொழிபெயர்ப்பு தொடர்பான பெரும்பாலான ஆராய்ச்சிகள் புள்ளிவிவர மற்றும் எடுத்துக்காட்டு அடிப்படையிலானவை நெருங்குகிறது. இயற்கை மொழி செயலாக்கத்தில் (என்.எல்.பி) இயந்திரமொழிபெயர்ப்பின் வகைப்பாட்டைக் கீழ்வரும் படம் காட்டுகிறது.

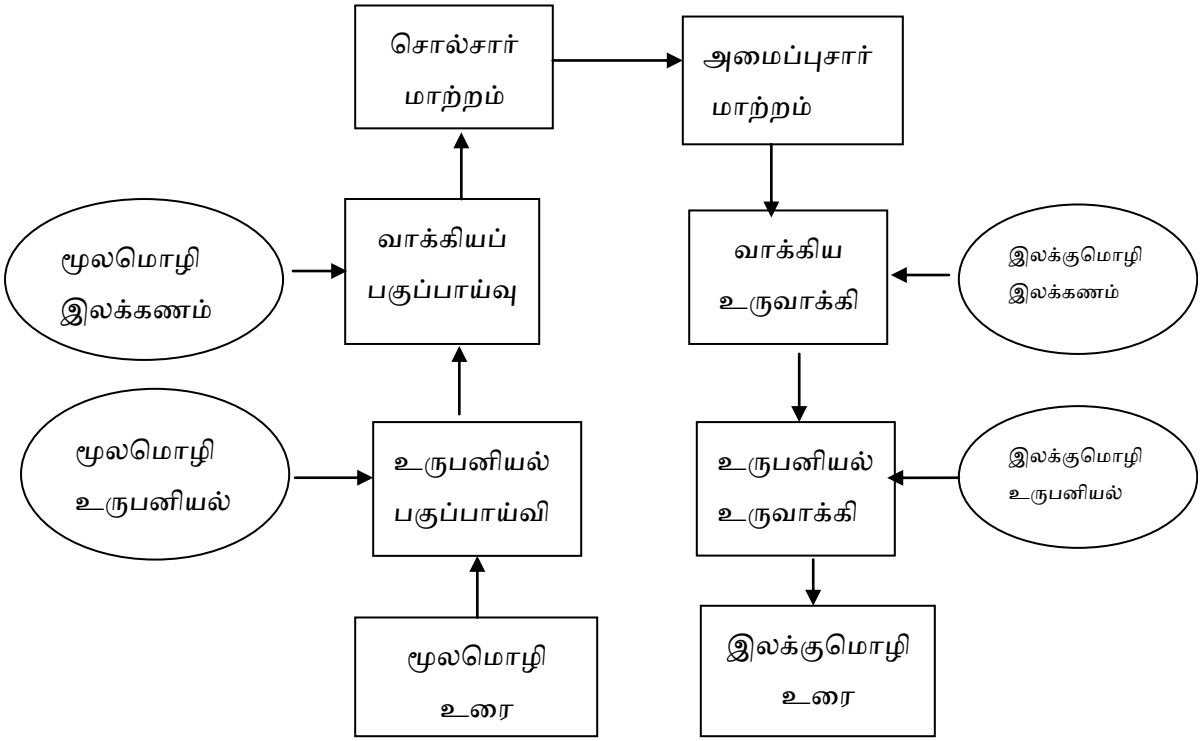


1.3.4.1. **விதி அடிப்படையிலான அணுகுமுறை**

ஒரு விதி அடிப்படையிலான ஒழுங்குமுறைக்கு மொழிபெயர்ப்பை அடைவதற்கு மூல மற்றும் இலக்கு மொழியைப் பற்றிய நிபுணர்களின் அறிவு தேவைப்படுகிறது, இது மொழிபெயர்ப்பை அடைய தொடரியல், பொருண்மையியல் மற்றும் உருபனியல் விதிகளை உருவாக்குகிறது. எடுத்துக்காட்டாக ஆங்கிலத்திலிருந்து ஜெர்மன் மொழிக்கு விதி அடிப்படையிலான மொழிபெயர்ப்புக்கு ஆங்கிலம்-ஜெர்மன் அகராதி தேவை, ஆங்கில இலக்கணத்திற்கான விதிகள் மற்றும் ஜெர்மன் இலக்கணத்திற்கான விதிகள் தேவை

ஒரு ஆர்.பி.எம்.டி ஒழுங்குமுறையில் சொற்கள் பிரிப்பு (Tokenisation/டோக்கனைசேஷன்), சொல்வகைப்பாடு அடையாளப்படுத்தல் மற்றும் பலவற்றை உள்ளடக்கிய இயற்கை மொழி

செயலாக்கம் (Natural Language Processing (NLP/என்.எல்.பி) பணிகளின் குழாய் (pipeline) உள்ளது. இந்த வேலைகளில் பெரும்பாலானவை மூல மற்றும் இலக்கு மொழியில் செய்யப்பட வேண்டும்.



இயந்திர மொழிபெயர்ப்பு துறையில் விதி அடிப்படையிலான அணுகுமுறை முதலில் உருவாக்கப்பட்ட உத்தி ஆகும். ஒரு விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Rule-Based Machine Translation (RBMT)) ஒழுங்குமுறை விதிகளின் தொகுப்பைக் கொண்டுள்ளது; இதில் இலக்கண விதிகள், இருமொழி அல்லது பன்மொழி அகராதி மற்றும் விதிகளை செயலாக்க மென்பொருள் நிரல்கள் என்பன அடங்கும். ஆயினும், RBMT ஒழுங்குமுறைகளை உருவாக்குவது மூல மொழி சொல்வகைபாட்டு அடையாளப்படுத்திகள் மற்றும் தொடரியல் பாகுபடுத்திகள், இருமொழி அகராதிகள், மூலமொழியிலிருந்து இலக்குமொழிக்கு எடுத்துப்பெயர்ப்பு, இலக்குமொழி உருபனியல் உருவாக்கி, கட்டமைப்பு மாற்றல், மற்றும் விதிகளை மறுவரிசைப்படுத்துதல் போன்ற மொழியியல் வளங்களை உட்படுத்தும். ஆயினும்கூட, ஒரு

RBMT அமைப்பு எப்போதும் நீட்டிக்கக்கூடியதும் பராமரிக்கக்கூடியதும் ஆகும். மொழிபெயர்ப்பின் பல்வேறு கட்டங்களில் விதிகள் முக்கிய பங்கு வகிக்கின்றன; அதாவது தொடரியல் செயலாக்கம், சொற்பொருள் விளக்கம் மற்றும் மொழியின் சூழ்நிலைச் செயலாக்கம் என்பன. பொதுவாக, மொழியியலாளர்களிடமிருந்து பெறப்பட்ட மொழியியல் அறிவைப் பயன்படுத்தி விதிகள் எழுதப்படுகின்றன. மாற்றல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு, இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மற்றும் அகராதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு என்பன மூன்று வெவ்வேறு அணுகுமுறைகள் RBMT வகையின் கீழ் வரும். ஆங்கிலத்திலிருந்து இந்திய மொழிகள் மற்றும் இந்திய மொழியிலிருந்து இந்திய மொழிகள் என்ற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் இந்த நான்கு அணுகுமுறைகளிலும் பலனளிக்கும். இவற்றின் பின்னணியில் உள்ள விதி அடிப்படையிலான அணுகுமுறைகள் முக்கிய கருத்துகள் பின்வருமாறு அமையும்.

நன்மைகள்

•இருமொழி உரை (bilingual text) தேவையில்லை

• பொருண்மைக்களம் சாராதது

• மொத்த கட்டுப்பாடு (ஒவ்வொரு சூழ்நிலைக்கும் சாத்தியமான புதிய விதி)

•மறுபயன்பாடு (புதிய மொழிகளுடன் இணையாக இருக்கும்போது இருக்கும் மொழிகளின் விதிகள் மாற்றப்படலாம்)

குறைபாடுகள்

• நல்ல அகராதிகள் தேவை

• கைமுறையாக விதிகளை அமைத்தல் (நிபுணத்துவம் தேவை)

• அதிக விதிகள் கணினியைக் கையாள்வது கடினம்

1.3.4.1.1.**நேரடி மொழிபெயர்ப்பு**

நேரடி மொழிபெயர்ப்பு முறையில் மூலமொழி உரை கட்டமைப்பு ரீதியாக உருபனியல் வரை பகுப்பாய்வு செய்யப்படுகிறது மற்றும் ஒரு குறிப்பிட்ட மூல மற்றும் இலக்கு மொழி இணைக்காக வடிவமைக்கப்பட்டுள்ளது (Noon et al 2003, Dasgupta & Basu 2008).). நேரடி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் செயல்திறன் மூல-இலக்கு மொழி அகராதிகளின் தரம் மற்றும் அளவு, உருபனியல் பகுப்பாய்வு, உரை செயலாக்க மென்பொருள் மற்றும் சொல்

வரிசையில் மற்றும் உருபனியலில் சிறிய இலக்கண மாற்றங்களுடன் சொல்லிருந்து சொல் மொழிபெயர்ப்பு இவற்றைச் சார்ந்து அமையும்.

1.3.4.1.2.**இடைமொழி அடிப்படையிலான மொழிபெயர்ப்பு**

இயந்திர மொழிபெயர்ப்பு அமைப்புகளின் வளர்ச்சியில் அடுத்த கட்ட முன்னேற்றம் இடைமொழி அடிப்படையிலான அணுகுமுறை ஆகும். இதில் மொழிபெயர்ப்பு மூலமொழி உரையை முதலில் இடைமொழி என்று அழைக்கப்படும் ஒரு இடைநிலை (சொற்பொருள்) வடிவில் உருப்படுத்தம் செய்வதன் மூலம் செய்யப்படுகிறது. இடைமொழி அணுகுமுறையின் நன்மை என்னவென்றால், மொழிசுதந்திரமான உருப்படுத்தத்திலிருந்து வேறுபட்ட இலக்குமொழிகளுக்கு மொழிபெயர்ப்புகளை உருவாக்க இயலும் என்பது ஆகும். இதனால், மொழிபெயர்ப்பு இரண்டு நிலைகளைக் கொண்டுள்ளது: இடைமொழியிலிருந்து இலக்கு மொழிக்கு மொழிபெயர்ப்பதற்கு முன் முதலில் மூலமொழி இடைமொழியாக (Interlingual Language (IL) மாற்றப்படுகிறது.. இந்த இடைமொழி அணுகுமுறையின் முக்கிய நன்மை என்னவென்றால், மூலமொழிக்கான பாகுபடுத்தியின்/பகுப்பானின் பகுப்பாய்வி இலக்குமொழி உருவாக்கியிலிருந்து சுதந்திரமாக உள்ளது. இடைமொழி அணுகுமுறையில் இரண்டு முக்கிய குறைபாடுகள் உள்ளன. முதல் தீமை என்னவென்றால், இடைமொழியை வரையறுப்பதில் சிரமம் உள்ளது. இரண்டாவது குறைபாடு இடைமொழி, மொழிகளுக்கு இடையிலான ஒற்றுமையின் நன்மையைக் கணக்கில் எடுக்கவில்லை; எடுத்துக்காட்டாக திராவிட மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பு போன்றவற்றில். ஆயினும்கூட இடைமொழியின் நன்மை என்னவென்றால், பல மொழிகளின் மொழிபெயர்ப்பில் ஈடுபட்டும் சூழ்நிலைகளில் இது சிக்கனமானது (Shachi et al, 2001).

கீழே உள்ள ஆழமற்ற மட்டத்திலிருந்து தொடங்கி, சொல் மட்டத்தில் நேரடி மாற்றல் செய்யப்படுகிறது. தொடரியல் மற்றும் பொருண்மையியல்சார் மாற்றல் அணுகுமுறைகள் மூலம் மேல்நோக்கி நகரும் போது, முறையே மூல வாக்கிய அமைப்பு மற்றும் பொருளின் உருப்படுத்தங்களில் மொழிபெயர்ப்பு நிகழ்கிறது. இறுதியாக இடைமொழி நிலையில் மாற்றலின் கருத்து 'இடைமொழி' என்று அழைக்கப்படு ஒரு அடிப்படை உருப்படுத்தத்தால் பதிலீடுசெய்யப்படுகிறது. 'இடைமொழி' ஒரே நேரத்தில் மூல மற்றும் இலக்குமொழி உரைகளை உருப்படுத்தம் செய்கின்றது. முக்கோணத்தை நகர்த்தினால் பயணிக்கத் தேவையான வேலையின்

அளவு குறையும் போது மொழிகளுக்கு இடையிலான இடைவெளியை நிரப்பத் தேவையான அளவு பகுப்பாய்வு மற்றும் உருவாக்கும் முயற்சி அதிகரிக்கும்.

1.3.4.1.3.**மாற்றல் அடிப்படையிலான மொழிபெயர்ப்பு**

இடைமொழி அணுகுமுறையின் குறைபாடு காரணமாக மாற்றல் அணுகுமுறை என்று அழைக்கப்படுகிற ஒரு சிறந்த விதி அடிப்படையிலான மொழிபெயர்ப்பு அணுகுமுறை கண்டுபிடிக்கப்பட்டது. மாற்றல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு என்பது ஒரு வகை இயந்திர மொழிபெயர்ப்பு ஆகும். இது தற்போது இயந்திர மொழிபெயர்ப்பின் மிகவும் பரவலாகப் பயன்படுத்தப்படும் முறைகளில் ஒன்றாகும். சமீபத்தில் வெளிநாடுகளிலும் இந்தியாவிலும் பல ஆராய்ச்சிக் குழுக்கள் தங்கள் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு இந்த மூன்றாவது அணுகுமுறையைப் பயன்படுத்துகின்றன.

மூலமொழிக்கும் இலக்கு மொழிக்கும் இடையிலான கட்டமைப்பு வேறுபாடுகள் அடிப்படையில் மாற்றல் அமைப்பு மூன்று வெவ்வேறு நிலைகளாக பிரிக்கப்பட்டுள்ளது: (1) பகுப்பாய்வு (analysis), (2) மாற்றல் (transfer) மற்றும் (3) உருவாக்கம் (generation). முதல் கட்டமாக மூலமொழி பாகுபடுத்தி/பகுப்பான் (parser) ஒரு மூலமொழி வாக்கியத்தின் தொடரியல் உருப்படுத்தத்தை உருவாக்கப் பயன்படுத்தப்படுகின்றது. அடுத்த கட்டத்தில் முதல் கட்டத்தின் விளைவு அதற்குச் சமமான இலக்குமொழி சார்ந்த உருப்படுத்தமாக மாற்றப்படுகிறது. இந்த மொழிபெயர்ப்பு அணுகுமுறையின் இறுதி கட்டத்தில், ஒரு இலக்குமொழி உருபனியல் உருவாக்கி இறுதி இலக்குமொழி உரைகளை உருவாக்கப் பயன்படுகிறது.

தற்கால ஒழுங்குமுறைகள் இந்த மூன்று அணுகுமுறைகளின் மேம்பாடுகளைப் பயன்படுத்துகின்றன. முழு ஒழுங்குமுறையும் பல எண்ணிக்கையிலான துணை ஒழுங்குமுறைகளாகப் பிரிக்கப்பட்டுச் சரியாக இடைமுகப்படுத்தப்பட்டுள்ளன. இயந்திர மொழிபெயர்ப்பின் இவ்வேறுபட்ட கட்டங்களை மூன்று நிலைகளாக் கூறலாம்.

1. ஆய்வு நிலை

ஒழுங்குமுறையின் உள்ளீடாக மூல மொழி உரை முதலில் சொல் நிலையில் ஆயப்படுகிறது. உருபனியல் ஆய்வு, உருபனியல் விதிகள், சந்தி விதிகள், அடிச்சொல், இயந்திரம் படிக்கவியலும் அகராதி மற்றும் முன்னொட்டு அகராதி போன்ற பலதரப்பட்ட அகராதிகளைப் பயன்படுத்திச் சொற்களை அவற்றின் உருபன்களுக்காகப் பிரிக்கின்றது. வாக்கியப் பகுப்பாய்வியில் ஒரு இலக்கண மாதிரியைப் பயன்படுத்தி ஒவ்வொரு வாக்கியத்தையும் அவற்றின் உறுப்புக்கான தொடர்கள் மற்றும் எச்சத் தொடர்களுக்காகப் பகுப்பாய்வு செய்கிறது; பயனிலைகளுக்கும் அவற்றின் பங்கெடுப்பாளர்களுக்கும் இடையில் உள்ள தொடரியல் மற்றும் பொருண்மையியல் உறவுகள் குறியாக்கம் செய்யப்படுகின்றன.

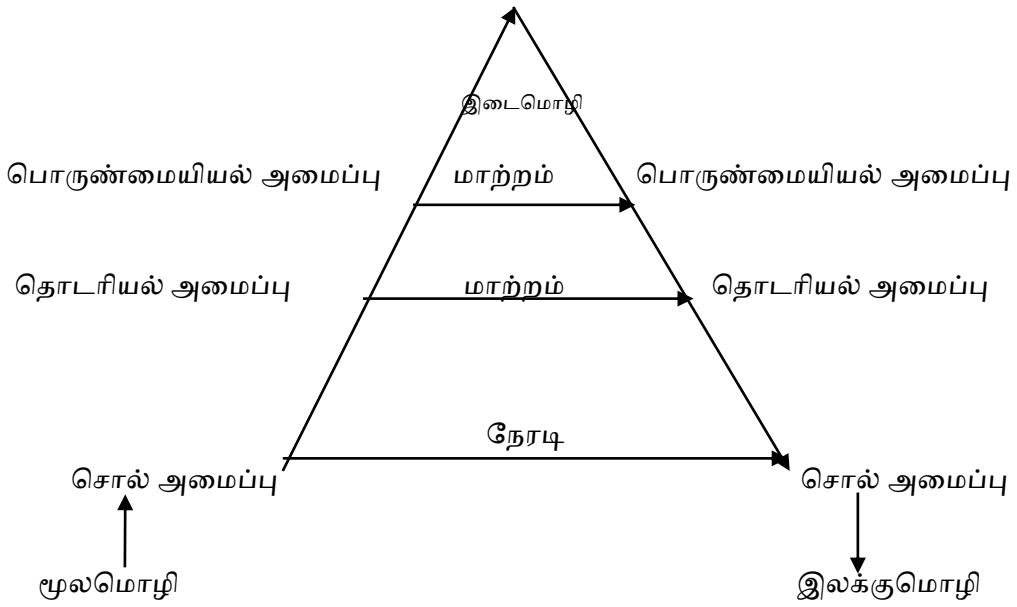
2. மாற்ற நிலை

இந்நிலையில் இரண்டு நிலைகள் உள்ளன: 1. சொல்சார் மாற்றம் 2. அமைப்புசார் மாற்றம். சொல்சார் மாற்ற நிலையில் உரை மற்றும் அகராதியின் தொடரியல் மற்றும் பொருண்மையியல் குறியாக்கத்தின் உதவியுடன் ஒவ்வொரு சொல்லின் சொல்சார் பொருண்மையின் மெய்ப்படுத்தம் செய்யப்படும்; இலக்கு மொழியிலிருந்து பொருத்தமான சொல் தேர்ந்தெடுக்கப்படும். அமைப்பு மாற்ற நிலையில் இலக்கு மொழி, மூல மொழி விதிகளைப் பயன்படுத்தி மூல மொழி அமைப்பு இலக்கு மொழி அமைப்பிற்கு மாற்றப்படும்.

3. உருவாக்க நிலை

இந்த நிலையில் வாக்கிய உருவாக்கும் பகுதி (Sentence Generator Module) இலக்கு மொழி இலக்கணத்தின் உதவியால் பெறப்படும் அமைப்புகளுக்கு இலக்கண அடிப்படையில் சரியான வாக்கியங்களை உருவாக்கும். பின்னர் உருபனியல் விதிகள் அல்லது புணர்ச்சி விதிகள், இயந்திரம் படிக்கவியலும் முன்னொட்டு அகராதி இவற்றைப் பயன்படுத்தி உருபனியல் சார் குறியாக்கம் இலக்கு மொழியின் பொருத்தமான ஒட்டுகளால் இடம் பெயர்க்கப்படும். வெளியீடு இலக்கு மொழியில் ஒரு உரையாக உற்பத்தி செய்யப்படும். இம்மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் வேறுபட்ட நிலைகளும் அவற்றின் துணை ஒழுங்குமுறைகளும் கீழே தரப்பட்டுள்ளன.

ஒரு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதில் ஏற்படும் சிக்கலின் அளவு மொழி பெயர்க்கப்படும் இரு மொழிகளின் ஒற்றுமை மற்றும் வேற்றுமைகளை பொறுத்து அமையும். இவ்வொற்றுமை வேற்றுமைகள் சொல் நிலையில் மொழி மாற்றம் செய்யப்பட வேண்டுமா? தொடர் நிலையில் மாற்றம் செய்யப்பட வேண்டுமா? பொருண்மை நிலையில் மாற்றம் செய்யப்பட வேண்டுமா என்பதைத் தீர்மானிக்கும். பெர்னார்ட் வாகோயிஸின் பிரமிடு (Bernard Vauquois' pyramid) இச்சாத்தியங்களை வெளிப்படுத்தும். பின்வரும் பெர்னார்ட் வாகோயிஸின் பிரமிடு (Bernard Vauquois' pyramid) நேரடி. இடைமொழி, மாற்றல் மொழிபெயர்ப்புகளின் உருப்படுத்தத்தின் ஒப்பீட்டு ஆழத்தைக் காட்டுகின்றது. இது இடைமொழ் இயந்திர மொழிபெயர்ப்பை உச்சத்தில் காட்டுகிறது, அதைத் தொடர்ந்து மாற்றல் அடிப்படையிலான மொழிபெயர்ப்பைக் காட்டுகின்றது; பின்னர் நேரடி மொழிபெயர்ப்பைக் காட்டுகின்றது.



நேரடி மொழிபெயர்ப்பின் போது மூலமொழிக்கும் இலக்குமொழிக்கும் உள்ள தூரம் அதிகமாக உள்ளதைக் காட்டுகின்றது. அடுத்ததாக மாற்றல் மொழிபெயர்ப்பின் போது அது தேர்வு செய்யும் நிலை (சொல்லமைப்பு, தொடரமைப்பு, பொருண்மையியல் அமைப்பு) அடிப்படையில் மூலமொழிக்கும் இலக்குமொழிக்கும் இடையிலான தூரம் குறைவதைக் காட்டுகின்றது. இடைமொழி நிலையில் மூலமொழிக்கும் இலக்குமொழிக்கும் இடையில் உள்ள தூரம் முற்றிலும் குறைவது தெரிகின்றது. இதன் மூலம் இம்மூன்றுவகையான மொழிபெயர்ப்பின் போது செய்யவேண்டிய முயற்சியின் வேறுபாட்டை ஊகிக்க இயலுகின்றது.

1.3.4.2. புள்ளியியல் அடிப்படையிலான அணுகுமுறை (Statisical approach)

இந்த அணுகுமுறை இருமொழி உரை நிறுவனத்தின் பகுப்பாய்வின் அடிப்படையில் புள்ளியியல் மாதிரிகளைப் பயன்படுத்துகிறது. இது முதன்முதலில் 1955இல் அறிமுகப்படுத்தப்பட்டது (Brown et al 1988, 1990). ஆனால் இது 1988க்குப் பிறகு ஐபிஎம் வாட்சன் ஆராய்ச்சி மையம் அதைப் பயன்படுத்தத் தொடங்கியபோதுதான் ஆர்வம் பெற்றது. புள்ளியியல்சார் இயந்திரமொழிபெயர்ப்புக்குப் பின்னால் உள்ள யோசனை பின்வருமாறு:

இலக்கு மொழியில் T என்ற வாக்கியத்தைக் கொடுத்தால், மொழிபெயர்ப்பாளர் T தயாரித்த S என்ற வாக்கியத்தை நாங்கள் தேடுகிறோம். அந்த வாக்கியத்தை S தேர்ந்தெடுப்பதன் மூலம் பிழையின் வாய்ப்பு குறைக்கப்படுவதை நாங்கள் அறிவோம். இது மிகவும் சாத்தியமான T. கொடுக்கப்பட்டுள்ளது. ஆகவே, எஸ் ஐ தேர்வு செய்ய விரும்புகிறோம் Pr (S | T) ஐ அதிகரிக்க.

- *A Statistical Approach to Machine Translation, 1990.*, (Brown 1988)

புள்ளியியல்சார் அணுகுமுறை அனுபவ இயந்திர மொழிபெயர்ப்பு (EMT) அமைப்புகளின் கீழ் வருகிறது; இது பெரிய இணையாக வரிசைப்படுத்தப்பட்ட தரவுத்தொகுதிகளை (lareg parallel aligned copora) நம்புகிறது. புள்ளிவிவர இயந்திர மொழிபெயர்ப்பு என்பது தரவு சார்ந்ததாகும். ஒரு இயற்கையான மொழியிலிருந்து இன்னொரு மொழிக்கு உரையை மொழிபெயர்ப்பதற்கான புள்ளியியல்சார் கட்டமைப்பு (stastical framework) இருமொழிய தரவுத்தொகுதியிலிருந்து (bilingual corpora) பிரித்தெடுக்கப்பட்ட அறிவு மற்றும் புள்ளிவிவர மாதிரிகள் (statistical models) அடிப்படையிலானதாகும். புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பில் மூல மற்றும் இலக்கு மொழி அல்லது மொழிகளின் இருமொழிய (bilingual) அல்லது பன்மொழிய (multilingual) உரைத் தரவுத்தொகுதி (text copora) தேவை. மேற்பார்வையிடப்பட்ட (supervised) அல்லது மேற்பார்வை செய்யப்படாத (unsupervised) புள்ளயியல்சார் இயந்திர கற்றல் வழிமுறை (machine learning algorithm) தரவுத்தொகுதிலிருந்து புள்ளியியல்சார் அட்டவணைகள் உருவாக்க பயன்படுகிறன; இந்தச் செயல்முறை கற்றல் (learning) அல்லது பயிற்சி (training) (Zhang et al 2006) என்று அழைக்கப்படுகின்றது' புள்ளிவிவர அட்டவணைகள் நன்கு உருவாக்கப்பட்ட வாக்கியங்களின் பண்புக்கூறுகள் மற்றும் மொழிகளுக்கு இடையிலான தொடர்பு போன்ற புள்ளிவிவர தகவல்களைக் கொண்டுள்ளன. மொழிபெயர்ப்பின் போது, சேகரிக்கப்பட்ட புள்ளிவிவர தகவல்கள் உள்ளீட்டு வாக்கியங்களுக்கான சிறந்த மொழிபெயர்ப்பைக் கண்டுபிடிக்கப் பயன்படுகின்றன; மற்றும் இந்த மொழிபெயர்ப்பு நடவடிக்கை குறியத்திறப்பு செயற்ப்பாங்கு (decoding process) என்று அழைக்கப்படுகிறது.

SMTக்குப் பின்னால் உள்ள கருத்து தகவல் கோட்பாட்டிலிருந்து வருகிறது. ஒரு ஆவணம் p(e|f) ஆல் குறிக்கப்பட்ட நிகழ்தகவு விநியோக செயல்பாட்டின் படி மொழிபெயர்க்கப்பட்டுகிறது;

இது மூலமொழி Fஇல் f ஒரு வாக்கியத்தை இலக்குமொழி Eஇல் e என்ற வாக்கியமாக மொழிபெயர்ப்பு செய்வதன் நிகழ்தகவு ஆகும்.

நிகழ்தகவு விநியோகம் p (e|f)ஐ மாதிரியாக்குவதின் சிக்கல் பல எண்ணிக்கையிலான வழிகளின் அணுகப்பட்டுள்ளது. ஒரு உள்ளுணர்வு அணுகுமுறை (intuitive approach) பேயஸ் தேற்றத்தைப் (Bayes theorem) பயன்படுத்துவதாகும். அதாவது, p (f|e) மற்றும் p(e) முறையே மொழிபெயர்ப்பு மாதிரி மற்றும் மொழி மாதிரியை குறிக்கிறது என்றால், பின்னர் நிகழ்தகவு விநியோகம் p(e|f) ∞ p(f|e)p(e) ஆகும். மொழிபெயர்ப்பு மாதிரி p(f|e) என்பது மூல வாக்கியத்தின் இலக்கு வாக்கியத்தின் மொழிபெயர்ப்பு அல்லது Eஇல் உள்ள வாக்கியங்கள் Fஇல் உள்ள வாக்கியங்களாக மாற்றப்படும் என்ற நிகழ்தகவு ஆகும். மொழி மாதிரி p(e) என்பது இலக்குமொழிக் கோர்வை அல்லது மொழி Eஇல் இருக்க இயலும் வாக்கியங்களின் வகைகளைக் காணும் நிகழ்தகவு ஆகும். இந்த சிதைவு சிக்கலை இரண்டு துணை சிக்கல்களாகப் பிரிக்கும்போது கவர்ச்சிகரமானதாக இருக்கிறது. சமன்பாடு 1 இல் காட்டப்பட்டுள்ளபடி சிறந்த மொழிபெயர்ப்பு ê-யைக் கண்டுபிடிப்பது அதிக நிகழ்தகவு தரும் ஒன்றைத் தேர்ந்தெடுப்பதன் மூலம் செய்யப்படுகிறது.

$$\hat{e} = \arg\max_{e \in e^*} p(e|f) = \arg\max_{e \in e^*} p(f|e)p(e)$$

சொற்றொடர் அடிப்படையிலான மாதிரிகள் SMTக்கு மிகவும் வெற்றிகரமான முறையாக வெளிப்பட்டிருந்தாலும் அவை தொடரியலை இயற்கையான முறையில் கையாளவில்லை. மொழிபெயர்ப்பின் போது சொற்றொடர்களை மறுவரிசைப்படுத்துவது பொதுவாக SMTஇல் விலகல் மாதிரிகளால் (distortion models) நிர்வகிக்கப்படுகிறது. ஆயினும்கூட, இந்த மறுவரிசைப்படுத்தும் செயற்பாங்கு (reordering process) முற்றிலும் திருப்தியற்றது, குறிப்பாக சொல் வரிசையின் அடிப்படையில் நிறைய வேறுபடுகிற மொழி இணகளுக்கு. மூலமொழிக்கும் இலக்குமொழிக்கும் உள்ள தொடரமைப்பு வேறுபாட்டுச் சிக்கலை அதிக அளவில் மறுவரிசைப்படுத்தும் செயல்பாட்டைச் செய்து வெற்றிகொள்ளலாம். உருபனியல் செழுமையான தமிழ் போன்ற மொழிகளுக்கு உருபனியல் தகவலைப் பயன்படுத்துவது பயிற்சி தரவு அளவைக் கணிசமாகக் குறைக்க இயலும் மற்றும் செயல்திறன் மேம்படும் (Rajendran and Vasuki, 2019) .

பேயஸின் தேற்றத்தைப் (Bayes' theorem) பயன்படுத்தி இந்த அதிகபட்ச சிக்கலை Pr (S) மற்றும் Pr (T | S) ஆகியவற்றின் பெருக்குத்தொகைக்கு மாற்றலாம், இங்கு Pr (S) என்பது S இன் மொழி மாதிரி நிகழ்தகவு (S என்பது அந்த இடத்தில் சரியான வாக்கியம்) மற்றும் Pr (T | S) என்பது T கொடுக்கப்பட்ட S இன் மொழிபெயர்ப்பு நிகழ்தகவு. வேறுவிதமாகக் கூறினால், ஒரு தேர்வுக்குரிய மொழிபெயர்ப்பு எவ்வளவு சரியானது மற்றும் சூழலில் அது எவ்வளவு பொருந்துகிறது என்பதைக் கொடுக்கும் மொழிபெயர்ப்பை நாங்கள் தேடுகிறோம்.

$$\Pr(S|T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)}$$

எனவே, ஒரு SMTக்கு மூன்று படிகள் தேவை: 1) ஒரு மொழி மாதிரி (Language Model) (அதன் சூழலில் கொடுக்கப்பட்ட சரியான சொல் என்ன?); 2) ஒரு மொழிபெயர்ப்பு மாதிரி (Translation Model) (கொடுக்கப்பட்ட வார்த்தையின் சிறந்த மொழிபெயர்ப்பு எது?); 3) சொற்களின் சரியான வரிசையைக் கண்டறிய ஒரு முறை.

முந்தைய பத்திகளில் வாக்கியம் மற்றும் சொல் இரண்டையும் மொழிபெயர்ப்பின் அலகுகளாகப் பயன்படுத்தினோம். அதிகம் பயன்படுத்தப்படும் மாதிரி இவற்றுக்கு இடையில் எங்காவது இருக்கும். இது தொடர் அடிப்படையிலான மொழிபெயர்ப்பு (phrase-based translation) என்று அழைக்கப்படுகிறது. எடுத்துக்காட்டாக, "is buying (வாங்குகிறது)" என்ற ஆங்கில சொற்றொடர் பிரெஞ்சு மொழியில் "achète" என்று மொழிபெயர்க்கப்பட்டுள்ளது.

நன்மைகள்

•இருமொழி உரை தேவையில்லை

•பொருண்மைக்களம் சாராதது

•மொத்த கட்டுப்பாடு (ஒவ்வொரு சூழ்நிலைக்கும் சாத்தியமான புதிய விதி)

•மறுபயன்பாடு (புதிய மொழிகளுடன் ஜோடியாக இருக்கும்போது இருக்கும் மொழிகளின் விதிகள் மாற்றப்படலாம்)

குறைபாடுகள்

•நல்ல அகராதிகள் தேவை

•கைமுறையாக விதிகளை அமைக்கவும் (நிபுணத்துவம் தேவை)

•கணினியைச் சமாளிப்பது கடினம்

மூன்று வெவ்வேறு புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்புகள் உள்ளன: சொல் அடிப்படையிலான மொழிபெயர்ப்பு, சொற்றொடர் அடிப்படையிலான மொழிபெயர்ப்பு மற்றும் படிநிலை தொடர் அடிப்படையிலான மாதிரி.

1.3.4.2.1. சொல் அடிப்படையிலான மொழிபெயர்ப்பு (word based translation)

இவ்வணுகுமுறையின் பெயர் குறிப்பிடுவது போல, உள்ளீட்டு வாக்கியத்தில் உள்ள சொற்கள் தனித்தனியாகச் சொற்களால் மொழிபெயர்க்கப்படுகின்றன; இறுதியாக இச்சொற்கள் இலக்கு மொழி வாக்கியத்தைப் பெற ஒரு குறிப்பிட்ட வழியில் வரிசைப்படுத்தப்பட்டுள்ளன. சொல் அடிப்படையிலான மொழிபெயர்ப்பில் உள்ளீடு மற்றும் வெளியீட்டு வாக்கியங்களில் உள்ள சொற்களுக்கு இடையில் வரிசைப்படுத்தல் பொதுவாக சில ஒழுங்கமைப்புகளைப் பின்பற்றுகிறது. இந்த அணுகுமுறை சொல் அடிப்படையிலான மொழிபெயர்ப்பில் முதல் முயற்சியாகும்; ஒப்பீட்டளவில் இது எளிமையான மற்றும் திறமையான புள்ளியியல்சார் அடிப்படையிலான இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையாகும். இந்த மொழிபெயர்ப்பு ஒழுங்குமுறையின் முக்கியமான குறைபாடு இதில் வாக்கியங்கள் சொல்லுக்குச் சொல் மொழிபெயர்க்கப்படுவதன் மூலம் மிகையாக எளிமைப்படுத்தப்பட்டுள்ளது; இதனால் மொழிபெயப்பு ஒழுங்குமுறையின் செயல்திறன் குறைகின்றது.

1.3.4.2.2. சொற்றொடர் அடிப்படையிலான மொழிபெயர்ப்பு

சொற்றொடர் அடிப்படையிலான மொழிபெயர்ப்பு Koehn et al, 2003) அறிமுகப்படுத்தப்பட்ட அணுகுமுறை ஒப்பீட்டளவில் மிகவும் துல்லியமான SMT ஆகும். இதில் மொழிபெயர்ப்புக்கு முன் ஒவ்வொரு மூல வாக்கியக்கியமும் இலக்குமொழி வாக்கியமும் தனித்தனி சொற்களாக அல்லாமல் சொற்றொடர்களாக பிரிக்கப்படுகின்றன. உள்ளீடு மற்றும் வெளியீட்டு வாக்கியங்களில் உள்ள சொற்றொடர்களுக்கு இடையிலான வரிசை முறை பொதுவாக சில ஒழுங்கமைப்புகளைப் பின்பற்றுகிறது; இது சொல் அடிப்படையிலான மொழிபெயர்ப்புடன் மிகவும் ஒத்திருக்கிறது. சொற்றொடர் அடிப்படையிலான மாதிரிகள் சொல் அடிப்படையிலான மொழிபெயர்ப்பை விட சிறந்த செயல்திறனை விளைவிக்கும் என்றாலும், அவை வாக்கிய வரிசை முறைகளின் மாதிரியை மேம்படுத்தவில்லை. வரிசைப்படுத்தும் மாதிரி மறுவரிசைப்படுத்தும் ஒழுங்கமைப்புகளை அடிப்படையாகக் கொண்டது மற்றும் சோதனைகள்

இந்த மறுவரிசைப்படுத்தும் நுட்பம் குறிப்பிட்ட இடம்சார் சொற்றொடர் வரிசைமுறைகளைக் கொண்டு செயல்படக்கூடும் என்பதைக் காட்டுகின்றன; ஆனால் நீண்ட வாக்கியங்கள் மற்றும் சிக்கலான வரிசைமுறைகளைக் கொண்டு செயல்பட இயலாது.

1.3.4.2.3. படிநிலை சொற்றொடர் அடிப்படையிலான மாதிரி (Hybrid Phrase model)

முந்தைய இரண்டு முறைகளின் குறைபாட்டைக் கருத்தில் கொண்டு, சியாங் (Chiang 2005) படிநிலை சொற்றொடர் அடிப்படையிலான மாதிரி (Hybrid Phrase model) என்று அழைக்கப்படுகிற மேலும் அதிநவீன SMT அணுகுமுறையையை உருவாக்கினார். இந்த அணுகுமுறையின் நன்மை என்னவென்றால், படிநிலை சொற்றொடர்கள் எளிய சொற்றொடர்களுக்குப் பதிலாக சுழல்நிலை/மறுசுழல் கட்டமைப்புகளைக் (recursive structures) கொண்டுள்ளன. இந்த உயர்நிலை சாராம்ச அணுகுமுறை SMT அமைப்பின் துல்லியத்தை மேலும் மேம்படுத்தியது.

1.3.4.2.4. கலப்பின அடிப்படையிலான மொழிபெயர்ப்பு (Hybrid-based approach)

புள்ளியியல் மற்றும் விதி அடிப்படையிலான மொழிபெயர்ப்பு முறைகளின் நன்மைகளைப் பெறுவதன் மூலம் கலப்பின அடிப்படையிலான அணுகுமுறை (Hybrid-based approach) என்று அழைக்கப்படுகிற புதியது அணுகுமுறை உருவாக்கப்பட்டது. இது இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகள் களத்தில் சிறந்தது என்று நிரூபிக்கப்பட்டுள்ளது. தற்போது, பல அரசு மற்றும் தனியார் சார்ந்த துறைகள் இயந்திரமொழிபெயர்ப்பு மூலமொழியிலிருந்து இலக்கு மொழிக்கு மொழிபெயர்ப்பை உருவாக்க இந்த கலப்பின அடிப்படையிலான அணுகுமுறையைப் பயன்படுத்துகின்றன; இது விதிகள் மற்றும் புள்ளிவிவரங்கள் இரண்டையும் அடிப்படையாகக் கொண்டது. கலப்பின அணுகுமுறையை பல் வெவ்வேறு வழிகளில் செயல்படுத்தலாம். சில சந்தர்ப்பங்களில் இந்த அணுகுமுறை விதிமுறைகளை அடிப்படையாகக் கொண்டு மொழிபெயர்ப்புகளை முதல் கட்டத்தில் செய்கின்றது; பின்னர் புள்ளியியல் தகவல்களைப் பயன்படுத்தி வெளியீட்டைச் சரிசெய்கின்றது. மற்றொரு வழியில் உள்ளீட்டுத் தரவை முன்செயலாக்கம் செய்யவும் புள்ளியியல் அடிப்படையிலான மொழிபெயர்ப்பின் வெளியீடை பின்செயலாக்கம் செய்யவும் விதிகள் பயன்படுத்தப்படுகின்றன. இந்த நுட்பம் முந்தையதை விடச் சிறந்தது மற்றும் மொழிபெயர்ப்பில் அதிக சக்தி, நெகிழ்வுத்தன்மை மற்றும் கட்டுப்பாடு உள்ளது.

ஒன்றுக்கு மேற்பட்ட இயந்திரமொழிபெயர்ப்பு முன்னுதாரணங்களை ஒருங்கிணைக்கும் கலப்பின அணுகுமுறைகளின் கவனம் அதிகரித்து வருகின்றன. METIS-II MT அமைப்பு EBMT கட்டமைப்பைச் சுற்றிய கலப்பினத்திற்கு ஒரு எடுத்துக்காட்டு ஆகும்; இது இருமொழி அகராதியைப் பயன்படுத்துவதன் மூலம் வழக்கமான இணையான தரவுத்தொகுதிகளின் மற்றும் இலக்குமொழியில் ஒரு ஒருமொழியத் தரவுத்தொகுதியின் தேவையைத் தவிர்க்கிறது (பெரும்பாலான RBMT அமைப்புகளில் காணப்படுவதைப் போன்றது) (Dirix et al., 2005). விதி அடிப்படையிலான முன்னுதாரணத்தைச் சுற்றியுள்ள கலப்பினத்தின் எடுத்துக்காட்டு ஓப்பனால் (Oepen) வழங்கப்படுகிறது. இது விதி அடிப்படையிலான நெறிமுறைகளைப் பயன்படுத்தி உருவாக்கப்படும் போட்டிக் கருதுகோள்களின் (மொழிபெயர்ப்புகள்) ஒரு தொகுப்பிலிருந்து சிறந்த மொழிபெயர்ப்பைத் தேர்வுசெய்ய RBMT ஒழுங்குமுறைக்குள் புள்ளிவிவர நெறிமுறைகளை ஒருங்கிணைக்கிறது (Oepen et al., 2007).

SMTஇல் கோஹன் மற்றும் ஹோங் ஆகியோர் உருபனியல், தொடரியல் அல்லது பொருண்மையியல் மட்டத்தில் விளக்கப்பட்ட மொழிபெயர்ப்பின் சில அம்சங்களைச் சிறப்பாகக் கற்றுக்கொள்வதற்காக மொழிபெயர்ப்பு மாதிரிகளுக்குள் சொல் மட்டத்தில் கூடுதல் சிறுகுறிப்புகளை (annotations) ஒருங்கிணைக்கின்றனர் (Koehn and et al., 2007). கலப்பின இயந்திரமொழிபெயர்ப்புக்கான புள்ளிவிவர அணுகுமுறையை க்ரோவ்ஸ் மற்றும் வே வழங்கியுள்ளனர்; அவர்கள் ஈபிஎம்டி ஒழுங்குமுறை மற்றும் எஸ்எம்டி ஒழுங்குமுறை இரண்டிலிருந்தும் சொற்றொடர்களை (துணை-அனுப்புதல்) சேர்த்துக்கொள்வதன் மூலம் தரவுத்தொகுதி அடிப்படையிலான முறைகள் இரண்டையும் ஒரு எஸ்எம்டி அமைப்பில் (Groves et al., 2005) இணைக்கின்றனர். ஒரு அடுக்கில் RBMT ஒழுங்குமுறையையும் SMT ஒழுங்குமுறையையும் பயன்படுத்தப்படும்போது ஒரு வித்தியாசமான கலப்பினமாக்கல் நிகழ்கிறது; சிமார்ட் (Simard) ஒரு RBMT அமைப்பால் உருவாக்கப்பட்ட மொழிபெயர்ப்புகளின் தானியங்கி பின் திருத்தியாகப் ஒரு SMT ஒழுங்குமுறையைப் பயன்படுத்தி டுகாஸ்ட்டுக்கு (Dugast) ஒப்பான ஒரு அணுகுமுறையை முன்மொழிந்தார் (Simard et al, 2007) (Dugast et al., 2007).

1.3.4.3. எடுத்துக்காட்டு அடிப்படையிலான மொழிபெயர்ப்பு (Example-based translation)

1984இல் மாகோடோ நாகோவால் முன்மொழியப்பட்ட எடுத்துக்காட்டு அடிப்படையிலான மொழிபெயர்ப்பு (Example-based translation (EBMT)அணுகுமுறை இரண்டு

மொழிபெயர்ப்பு எடுத்துக்காட்டுகளுக்கும் இடையிலான ஒப்புமை பகுத்தறிவை (analogical reasoning) அடிப்படையாகக் கொண்டது. இயங்கும் நேரத்தில் ஒரு எடுத்துக்காட்டு அடிப்படையிலான மொழிபெயர்ப்பு இருமொழியத் தரவுத்தொகுதியை அதன் முக்கிய அறிவுத் தளமாகப் பயன்படுத்துவதன் மூலம் பண்பாக்கம்செய்யப்படுகிறது. எடுத்துக்காட்டு அடிப்படையிலான அணுகுமுறை EMT அமைப்பின் கீழ் வருகிறது; இது பெரிய இணையாக வரிசைப்படுத்தப்பட்ட தரவுத்தொகுதிகளை நம்பியுள்ளது.

எடுத்துக்காட்டு அடிப்படையிலான மொழிபெயர்ப்பு அடிப்படையில் ஒப்புமை (analogy) மூலமான மொழிபெயர்ப்பாகும். ஒரு ஈபிஎம்டி (EBMT) ஒழுங்குமுறை மூலமொழி வாக்கியங்களின் தொகுப்பும் (அதிலிருந்து ஒருவர் மொழிபெயர்க்கிறர்) அவற்றுடன் தொடர்புடைய இலக்குமொழி மொழிபெயர்ப்புகளும் கொடுக்கப்பட்டுள்ளன; மற்றும் இந்த எடுத்துக்காட்டுகளை அதை ஒத்த பிற மூல-மொழி வாக்கியங்களை இலக்குமொழியில் மொழிபெயர்க்க பயன்படுத்துகின்றது. அடிப்படை முதற்கொள், முன்னர் மொழிபெயர்க்கப்பட்ட வாக்கியம் ஏற்பட்டால் மீண்டும் அதே மொழிபெயர்ப்பு மீண்டும் சரியாக இருக்கும். ஈபிஎம்டி ஒழுங்குமுறைகள் கவர்ச்சிகரமானவை; அவற்றிற்கு குறைந்தபட்ச முன் அறிவு தேவை; எனவே, அவை விரைவாக பல மொழி இணைகளுக்கும் ஏற்றதாக இருக்கும்.

எடுத்துக்காட்டு அடிப்படையிலான மொழிபெயர்ப்பின் கட்டுபடுத்தப்பட்ட வடிவம் வணிக ரீதியாக கிடைக்கிறது; இது மொழிபெயர்ப்பு நினைவகம் என அழைக்கப்படுகிறது. பயனர் உரையை மொழிபெயர்க்கும் போது அந்த மொழிபெயர்ப்புகளும் மொழிபெயர்ப்பு நினைவகத்தில் உள்ளன தரவுத்தளத்தில் சேர்க்கப்படும்; அதே வாக்கியம் மீண்டும் நிகழும்போது, முந்தைய மொழிபெயர்ப்பு மொழிபெயர்க்கப்பட்ட ஆவணத்தில் செருகப்படும். இது அந்த வாக்கியத்தை மீண்டும் மொழிபெயர்க்கும் முயற்சியைப் பயனருக்கு இலாபப்படுத்துகின்றது வாக்கியம்; மற்றும் குறிப்பாக முன்பு மொழிபெயர்க்கப்பட்ட ஆவணத்தின் ஒரு புதிய பதிப்பை மொழிபெயர்க்கும்போது பயனுள்ளதாக இருக்கும்.

மேலும் மேம்பட்ட மொழிபெயர்ப்பு நினைவக ஒழுங்குமுறைகளும் நெருங்கிய நிகரனின் மொழிபெயர்ப்பைத் திருத்துவதற்கு எடுக்கும் நேரம் புதிதாக ஒரு மொழிபெயர்ப்பை உருவாக்குவதற்கு எடுக்கும் நேரத்தைவிட குறைவான நேரம் எடுக்கும் என்ற அனுமானம் நெருக்கமான ஆனால் துல்லியமான நிகரன்களைத் தரும். ALEPH, wEBMT, ஆங்கிலத்திலிருந்து

துருக்கிய மொழி, ஆங்கிலத்திலிருந்து ஜப்பானிய மொழி, ஆங்கிலத்திலிருந்து சமஸ்கிருதம் மற்றும் PanEBMT ஆகியவை எடுத்துக்காட்டு அடிப்படையிலான எம்டி அமைப்புகள் ஆகும்.

பின்வரும் ஆங்கிலம்-ஜப்பானீஸ் இருமொழி தரவுத்தொகுதியின் எடுத்துக்காட்டைக் கருத்தில் கொள்ளவும்.

ஆங்கிலம்	ஜப்பனீஸ்
How much is that red umbrella?	Ano akai kasa wa ikura desu ka.
How much is that small camera?	Ano chiisai kamera wa ikura desu ka.

எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மேலே உள்ள அட்டவணையில் காட்டப்பட்டுள்ள எடுத்துக்காட்டு போன்ற வாக்கிய இணைகளைக் கொண்ட இருமொழி இணையான தரவுத்தொதியிலிருந்து பயிற்சி அளிக்கப்படுகின்றன. வாக்கிய இணைகளில் ஒரு மொழியில் உள்ள வாக்கியங்களும் மற்றொரு மொழியில் அவற்றின் மொழிபெயர்ப்புகள் உள்ளன. குறிப்பிட்ட எடுத்துக்காட்டு குறைந்தபட்ச இணையை எடுத்துக் காட்டுகிறது; அதாவது வாக்கியங்கள் ஒரு உறுப்பு (அல்லது சொல்) மூலம் வேறுபடுகின்றதைக் காட்டுகின்றது. இந்த வாக்கியங்கள் ஒரு வாக்கியத்தின் பகுதிகளின் மொழிபெயர்ப்புகளைக் கற்றுக்கொள்வதை எளிதாக்குகின்றன. எடுத்துக்காட்டாக, எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு முறை மேலே உள்ள எடுத்துக்காட்டில் இருந்து மூன்று உறுப்புகளின் மொழிபெயர்ப்பைக் கற்றுக் கொள்ளும்

How much is that X ? corresponds to Ano X wa ikura desu ka.

red umbrella corresponds to akai kasa

small camera corresponds to chiisai kamera

இந்த உறுப்புகளை உருவாக்குவது எதிர்காலத்தில் புதிய மொழிபெயர்ப்புகளை உருவாக்க பயன்படுகிறது. எடுத்துக்காட்டாக, சில வாக்கியங்களைக் கொண்ட உரையைப் பயன்படுத்தி நமக்குப் பயிற்சி அளிக்கப்பட்டிருந்தால்:

President Kennedy was shot dead during the parade. and The convict escaped on July 15th.

நாம் The convict was shot dead during the parade என்ற வாக்கியத்தை வாக்கியத்தின் பொருத்தமான பகுதிகளால் இடம்பெயர்த்து மொழிபெயர்க்கலாம்.

தொடர் வினைச்சொற்கள் (phrasal) போன்ற துணை மொழி நிகழ்வுகளுக்கு எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மிகவும் பொருத்தமானது. தொடர் வினைச்சொற்கள் மிகவும் சூழல் சார்ந்த அர்த்தங்களைக் கொண்டுள்ளன. அவை ஆங்கிலத்தில் பொதுவானவை; அங்கு அவை ஒரு வினைச்சொல்லைத் தொடர்ந்து ஒரு வினையடை மற்றும் / அல்லது ஒரு முன்னுருபைக் (preposition) கொண்டிருக்கின்றன, அவை வினைச்சொல்லின் இடைச்சொல் (particle) என்று அழைக்கப்படுகின்றன. தொடர் வினைச்சொற்கள் சிறப்பு சூழல் சார்ந்த அர்த்தங்களை உருவாக்குகின்றன; அவை உறுப்புகளின் பொருளிலிருந்து பெறப்படாமல் இருக்கலாம். மூலமொழியிலிருந்து இலக்கு மொழிக்கு சொல்லுக்குச் சொல் மொழிபெயர்ப்பின் போது எப்போதும் ஒரு பொருண்மைமயக்கம் காணப்படும். எடுத்துக்காட்டாக, put on "போடு" என்ற சொற்றொடர் வினையின் உருது/இந்தி பொருளைக் கவனியுங்கள். இது பின்வரும் வழிகளில் பயன்படுத்தப்படலாம்:

Ram put on the lights. (Switched on) (Urdu/Hindi translation: jalana)

Ram put on a cap. (Wear) (Urdu/Hindi translation: pahenna)

1.3.4.4. அறிவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Knowledge-based MT)

அறிவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Knowledge-based MT (KBMT)) அதிக முக்கியத்துவம் வாய்ந்ததாக வகைப்படுத்தப்படுகிறது (Nirenburg et al 1986). அறிவு அடிப்படையிலான இயந்திர மூலமொழி உரையை இலக்குமொழி உரையாக மொழிபெயர்ப்பதற்கு முன்னர் செயல்பாட்டு அடிப்படையில் மூலமொழி உரையின் முழுமையான புரிதல் மீதான ஆழுத்தத்தல் பண்பாக்கம் செய்யப்பட்டுள்ளது. KBMTக்கு மொத்த புரிதல் தேவையில்லை; ஆனால் பொருள்கோள் இயந்திரம் (interpretation engine) பல மொழிகளில் வெற்றிகரமான மொழிபெயர்ப்பை அடைய முடியும் என்று கருதுகிறது. கெபிஎம்டி (KBMT) இடைமொழிய கட்டமைப்பில் செயல்படுத்தப்படுகிறது; இது மூலம்மொழியை ஆழமாகப் பகுப்பாய்வு செய்வதாலும் உலகின் வெளிப்படையான அறிவை நம்பியிருப்பதாலும் இடைமொழி நுட்பங்களிலிருந்து வேறுபடுகிறது.

கெபிஎம்டி (KBMT) உலக அறிவாலும் சொற்களின் அர்த்தங்கள் மற்றும் அவற்றின் சேர்க்கைகள் பற்றிய மொழியியல் பொருண்மையியல்சார் அறிவாலும் ஆதரிக்கப்பட வேண்டும். எனவே ஒரு குறிப்பிட்ட மொழி, மொழிகளின் பொருளைக் உருப்படுத்தம் செய்யத் தேவை ஆகும்.

மூலமொழி பகுப்பாய்வு செய்யப்பட்டவுடன் மேம்படுத்தி வழி அது இயங்கும். அறிவுத் தளமே (knowledge) இலக்குமொழி வாக்கியத்தை உருவாக்குவதற்கு முன் மூலமொழி உருப்படுத்ததைப் பொருத்தமான இலக்குமொழி உருப்படுத்தமாதாக மாற்றுகிறது..

கெபிஎம்டி (KBMT) ஒழுங்குமுறைகள் உயர்தரமான மொழிபெயர்ப்புகளை வழங்குகின்றன. ஆயினும்கூட, வேறுபட்ட மொழிகளில் வாக்கியங்களை துல்லியமாக உருப்படுத்தம் செய்ய தேவையான பெரிய அளவிலான அறிவு காரணமாக அவை மிகவும் விலை உயர்ந்தவை ஆகும். ஆங்கிலம்-வியட்நாமிய இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை KBMTSஇன் எடுத்துக்காட்டுகளில் ஒன்றாகும்.

1.3.4.5. கோட்பாடு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Principle-Based MT)

கோட்பாடு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (பிபிஎம்டி (Principle-Based MT PBMT)) சாம்ஸ்கியின் ஆக்கமுறை இலக்கணத்தின் கோட்பாடுகள் மற்றும் அளவுருக்கள் அடிப்படையில் பாகுபடுத்தும் முறைகளைப் பயன்படுத்துகின்றன. பாகுபடுத்தி சொல்சார், இலக்கணம்சார் மற்றும் கருப்பொருள்சார் தகவல்கள் ஆகியவற்றைக் கொண்டுள்ள ஒரு விரிவான தொடரியல் அமைப்பை உருவாக்குகிறது. இது வலுவான தன்மை, மொழி-நடுநிலை உருப்படுத்தங்கள் மற்றும் ஆழமான மொழியியல் பகுப்பாய்விலும் கவனம் செலுத்துகிறது.

பிபிஎம்டியில் இலக்கணம் மொழி-சுதந்திரமான, ஊடாடும் நன்கு வடிவமைக்கப்பட்ட கொள்கைகள் மற்றும் மொழி சார்ந்த அளவுருக்களின் (language-dependent parameters) கணம் மற்றும் கொள்கைகளின் தொகுதி இவற்றின் ஒரு தொகுப்பாகக் கருதப்படுகிறது. இவ்வாறு, n மொழிகளைப் பயன்படுத்துகிற ஒரு ஒழுங்குமுறைக்கு, ஒருவர் n அளவுருக்களின் தொகுதிகளைக் கொண்டிருருக்க வேண்டும். இவ்வாறு, இது இடைமொழிய கட்டமைப்புக்குப் பயன்படுத்த மிகவும் பொருத்தமானது.

பிபிஎம்டி பாகுபடுத்தும் முறைகள் விதி அடிப்படையிலான அணுகுமுறைகளிலிருந்து வேறுபடுகின்றன. பல சூழ்நிலைகளில் திறமையாக இருந்தாலும், அவை மொழி சார்ந்திருத்தல் மற்றும் ஒருவர் பன்மொழிய மொழிபெயர்ப்பு முறையைப் பயன்படுத்தினால் பன்மடங்குகளில் விதிகளில் அதிகரிப்பு ஆகியவற்றின் குறைபாட்டைக் கொண்டுள்ளன. அவை பரந்த அளவில் பல மொழியியல் நிகழ்வுகளின் விரிந்த செயலெல்லையை/பரப்பெல்லையை வழங்குகின்றன; ஆனால் கெபிஎம்டி (KBMT) மற்றும் ஈபிஎம்டி (EBMT) அமைப்புகள் பயன்படுத்தும் மொழிபெயர்ப்பு களம்

பற்றிய ஆழமான அறிவு இல்லை. தற்போதைய பிபிஎம்டி அமைப்புகளின் மற்றொரு குறைபாடு வெவ்வேறு கொள்கைகளைப் பயன்படுத்துவதற்கான மிகவும் திறமையான முறையின் பற்றாக்குறை. யுனிட்ரான் (UNITRAN), பிபிஎம்டியின் எடுத்துக்காட்டுகளில் ஒன்று.

1.3.4.6. இணைநிலை/ ஆன்லைன் ஊடாடும் ஒழுங்குமுறைகள் (Online Interactive Systems)

இந்த ஊடாடும் மொழிபெயர்ப்பு ஒழுங்குமுறையில் மொழிபெயர்ப்பாளர் இணைநிலையில்/ஆன்லைனில் சரியான மொழிபெயர்ப்பை பரிந்துரைக்க பயனர் அனுமதிக்கப்படுகிறார். ஒரு வார்த்தையின் சூழல் தெளிவற்று இருக்கும் போதும் பயன்படுத்தும் சூழ்நிலையில் ஒரு குறிப்பிட்ட வார்த்தைக்குப் பல சாத்தியமான அர்த்தங்கள் இருக்கும் சந்தர்ப்பங்களிலும் இந்த அணுகுமுறை மிகவும் பயனுள்ளதாக இருக்கும். கட்டமைப்புசார்/அமைப்புசார் பொருண்மைமயக்கம் (structural ambiguity) பயனரின் சரியான பொருள்கோளினால் தீர்க்கப்படலாம்.

1.3.4.7. நரம்பியல் இயந்திர மொழிபெயர்ப்பு (Neural Machine Translation (NMT)

இயந்திர மொழிபெயர்ப்பை அடைய நரம்பியல் அணுகுமுறை நரம்பியல் வலைப்பின்னல்களை (networks/நெட்வொர்க்குகள்) பயன்படுத்துகிறது. முந்தைய மாதிரிகளுடன் (models/மாடல்கள்) ஒப்பிடும்போது, என்எம்டிகளை தனித்தனி பணிகளின் குழாய்வழிக்கு பதிலாக ஒரு பிணையத்துடன் உருவாக்க முடியும்.

2014ஆம் ஆண்டில், இயற்கைமொழி ஆய்வில் நரம்பியல் வலைப்பினனல்களுக்கு (networks/ நெட்வொர்க்குகள்) புதிய சாத்தியங்களைத் திறக்கும் வரிசை-க்கு-வரிசை மாதிரிகள் அறிமுகப்படுத்தப்பட்டன. Seq2seq மாதிரிகளுக்கு முன், நரம்பியல் வலைப்பினனல்கள் வரிசை உள்ளீட்டை கணினி-தயார் எண்களாக மாற்ற ஒரு வழி தேவைப்பட்டது [(ஒரு-சூடான குறியாக்கம் (one-hot encoding), உட்பொதிப்புகள் (embeddings)]. Seq2seq உடன், உள்ளீடு மற்றும் வெளியீட்டு வரிசைகளுடன் ஒரு பிணையத்தைப் பயிற்றுவிப்பதற்கான சாத்தியம் சாத்தியமானது (Sutskeve et al, Cho, et al).

சில வருட ஆராய்ச்சிக்குப் பிறகு என்எம்டி விரைவாக வெளிப்பட்டது. இந்த மாதிரிகள் எஸ்எம்டிகளை (SMT)விடச் சிறப்பாக செயல்பட்டன (Bahdanau et al 2014). மேம்பட்ட முடிவுகளுடன், பல மொழிபெயர்ப்பாளர் வழங்குநர் நிறுவனங்கள் தங்கள் வலைப்பின்னலகளை

(Networks/நெட்வொர்க்குள்) கூகிள் (Wu et al 2016) மற்றும் மைக்ரோசாப்ட் உள்ளிட்ட நரம்பியல் அடிப்படையிலான மாடல்களாக மாற்றின.

பயிற்சி தரவு சமநிலையற்றதாக இருந்தால் நரம்பியல் வலைப்பினனல்களில் சிக்கல் ஏற்படுகிறது. அரிய மாதிரிகள் மற்றும் அடிக்கடி வரும் மாதிரிகளிலிருந்து மாதிரியைக் கற்றுக்கொள்ள முடியாது. மொழிகளைப் பொறுத்தவரை, இது ஒரு பொதுவான பிரச்சினையாகும்; ஏனெனில் முழு விக்கிபீடியாவிலும் பல அரிய சொற்கள் சில முறை மட்டுமே பயன்படுத்தப்படுகின்றன. தொடர்ச்சியான சொற்களுக்கு பக்கச்சார்பற்ற ஒரு மாதிரியைப் பயிற்றுவிப்பது (எ.கா.: ஒவ்வொரு விக்கிபீடியா பக்கங்களிலும் பல நிகழ்வுகள்) சவாலானது. இந்த அரிய சொற்களை ஒரு அகராதி மூலம் மொழிபெயர்க்க ஒரு பிந்தைய செயலாக்க நடவடிக்கையைப் பயன்படுத்தி ஒரு சமீபத்திய கட்டுரை முன்மொழிகிறது (Luong et al 2014).

சமீபத்தில், பேஸ்புக் ஆராய்ச்சியாளர்கள் எஸ்எம்டி மற்றும் என்எம்டி இரண்டிலும் பணிபுரியும் மேற்பார்வை செய்யப்படாத இயந்திரமொழிபெயர்ப்பு மாதிரியை அறிமுகப்படுத்தினர், இதற்கு பெரிய ஒருமொழி தரவுத்தொகுதி மட்டுமே தேவைப்படுகிறது; இருமொழி தரவுத்தொகுதி அல்ல (Lample et al 2018). முந்தைய எடுத்துக்காட்டுகளின் முக்கிய சிக்கல் என்னவென்றால், பயிற்சியளிக்க மொழிபெயர்ப்புகளுடன் பெரிய தரவுத்தளம் இல்லாதது. இம்மாதிரி இச்சிக்கலை தீர்க்கும் உறுதிமொழியைக் காட்டுகிறது.

என்எம்டி எடுத்துக்காடுகள் (NMT examples)

- Google Translate (from 2016) link to language team at Google AI
- Microsoft Translate (from 2016) link to MT research at Microsoft
- Translation on Facebook: link to NLP at Facebook AI
- OpenNMT: An open-source neural machine translation system. [16]

நன்மைகள்

• முடிவுக்கு இறுதி மாதிரிகள் (End-to-end models) (குறிப்பிட்ட பணிகளின் குழாய் இல்லை)

குறைபாடுகள்

• இருமொழி தரவுத்தொகுதிகள் (bilingual corpus) தேவை

• அரிதான சொல் (Rare word) சிக்கல்

1.4. இயந்திரமொழிபெயர்ப்புக்கான இலக்கண வடிவமைப்புகள்

பலவிதமான இலக்கண வடிவமைப்புகள் (grammatical formats) மொழிபெயர்ப்பிற்குப் பயன்படுத்தப்படுகின்றன. எடுத்துக்காட்டாக முற்றுநிலை தானியங்கி (Finite State Automata

(FSA)), சூழல் கட்டிலா இலக்கணம் (Context Free Grammar (CFG)) , சூழல் கட்டுண்ட இலக்கணம் (Context Sensitive Grammar (CSG)), மாற்றிலக்கணம் (Transformational Grammar), கிளைசேர்ப்பு இலக்கணம் (Tree Adjoining Grammar (TAG)), தொடரமைப்பு இலக்கணம் (Phrase Structure Grammar (PSG)) , தலை இயக்கத் தொடரமைப்பு இலக்கணம் (Head Driven Phrase Structure Grammar (HDPSG), பாணினி இலக்கணம் (Panini Grammar (PG)), பொதுமைபடுத்தப்பட்ட தொடரமைப்பு இலக்கணம் (Generalized Phrase Structure Grammar (GPSG)) என பல இலக்கண வடிவமைப்புகள் மொழி ஆய்விற்காகப் பயன்படுத்தப்படுகின்றன. ஆளுகைக் கட்டுறவுக் கோட்பாடு (Government and Binding (GB) Theory) , சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)), கிளை இணைக்கும் இலக்கணம் (Tree Adjoining Grammar (TAG)), பெரிதாக்கப்பட்ட மாற்ற வலைப்பின்னல் (Augmented Transition Network (ATN)) போன்ற புதிய இலக்கண வடிவங்களின் வருகை இயற்கை மொழி ஆய்வுப் பரப்பில் புதிய பார்வையை உருவாக்கியது. இம்மொழியியல் வடிவமைப்புகள் பகுத்தாயும் வழிமுறைகளில் தொடரியல் மற்றும் பொருண்மையியல் பண்புக்கூறுகளின் பொருத்தத்திற்கு ஆலோசனைக்குரியது. மேலும் கணிப்பொறி அறிவியலில் ஏற்பட்ட தொழில்நுட்ப முன்னேற்றம் இயந்திர மொழிபெயர்ப்பு ஆய்வின் வளர்ச்சிக்குச் சாதமாக அமைந்தது. சொல்-பொருண்மையியல் பகுப்பாய்வை உள்ளடக்கிய இயற்கை மொழியின் புரிதல் (Natural Language Understanding) தேவையான உலக அறிவை அடையாளம் காணுவதை உள்ளடக்கியது. அறிவு உருப்படுத்தம் மற்றும் மொழியியல் அறிவில் உள்ளடக்குவது போன்றவை இயந்திர மொழிபெயர்ப்பு ஆய்வுகளின் பரப்பெல்லையில் மேலும் கூடுதல் வளர்ச்சிக்கு உதவி புரிந்தது.

1.5. இயந்திர மொழிபெயர்ப்புக்கான முன் நடவடிக்கைகள்

இயந்திர மொழிபெயர்ப்புக்கு மொழியியல் ஆய்வு மிக முக்கியமான ஒன்றாகக் கருதப்படுகிறது. மொழியியல் ஆய்வு வாக்கியங்களைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை உருபன்களாகவும் உருபன்களை ஒலியன்களாகவும் பகுத்தாய்கிறது. இவ்வாறு ஒரு மொழியானது அதன் கலவைத் தன்மையான அமைப்பை மெய்ப்படுத்தம் செய்ய மேற்சொன்ன படிநிலைகளைக் கொண்டிருக்கின்றன. இப்படிநிலைகள் தொடரியல், உருபனியல், ஒலியனியல் என்ற மொழி நிலைகளை உருப்படுத்தம் செய்கின்றன. வாக்கியங்களிலிருந்து ஒலியன் வரையிலான பகுப்பாய்வு எல்லாமே பொருண்மையைப் பெறுவதற்காக மனித மூளை

செய்யும் செயல்பாடுகளின் மாதிரி வடிவங்களாக மொழியியல் கோட்பாடுகள் எடுத்துக்கொள்ளகின்றன.

கணினி மொழியியலார் கணினியைப் பயன்படுத்தி மொழிக்கூறுகளை ஒலியன் நிலை, உருபன் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை என்ற நிலைகளில் ஆராய்ந்து இயந்திர மொழிபெயர்ப்புக்கு வேண்டிய தவல்களைத் தருகின்றனர். உருபனியல் பகுப்பாய்வு, சொல்வகைப்பாடு அடையாளப்படுத்தல், தொடர் உறுப்பு அடையாளப்படுத்தல், எச்சத் தொடர்களை அடையாளம் காணுதல் ஆகியன இயந்திர மொழிபெயர்ப்பு அமைப்பொழுங்கின் பகுதிகளாகும். மூல மொழியை இலக்கு மொழியாக மாற்ற அடிப்படையான செயல்பாடுகள் இவையாகும். இச்செயல்பாடுகள் மூலமொழி வாக்கியங்களின் பொருண்மையை வெளிப்படுத்தும் செயல்பாட்டை நோக்கமாகக் கொண்டதாகும். இத்தகைய நோக்கம் தழுவிய செயல்பாடுகளின் விளைவாகச் சொல் பிரிப்பான் (tokenizer), உருபனியல் ஆய்வி (morphological analyzer), சொல்வகைப்பாடு அடையாளப்படுத்தி (POS tagger), தொடரியல் பகுப்பான் (syntactic Parser) அல்லது தொடர்கூறு பகுப்பான் (chunker), உருபனியல் உருவக்கி (morphological generator) அல்லது சொல் உருவாக்கி (word generator), வாக்கிய உருவக்கி (sentece generator) ஆகியன உருவாக்கப்படுகின்றன.

தமிழ் போன்ற உருபனியல் அடிப்படையில் வளமுள்ள மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கப்பட வேண்டுமானால் உருபனியல் பகுப்பாய்வு செம்மையாகச் செய்யப்பட வேண்டும் என்பதைக் கருத்தில் கொண்டு உருபனியல் பகுப்பாய்விகள் (Morphological Analyzers) உருவாக்கப்படுகின்றன (Rajendran, Ganesan, Deyvasundaram, Vaishnavi). உருபனியல் பகுப்பாய்வி இயந்திர மொழிபெயர்ப்பின் முக்கியமான ஒரு பாகமாகக் கருதப்பட்டு அதை உருவாக்கப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன. ஆங்கிலம் போன்ற மொழிகளுக்கு உருபனியல் வளத்தைக் காட்டிலும் தொடரியல் வளம் கூடுதலாக இருந்ததன் காரணமாகத் தொடரியல் பகுப்பாய்விற்கு முக்கியத்துவம் தரப்பட்டுத் தொடரியல் பகுப்பான்கள் (Syntactic Parsers) உருவாக்கப்பட்டன.

உருபனியல் பகுப்பாய்வும் தொடரியல் பகுப்பாய்வும் மூல மொழியில் செய்யப்பட்டு அதில் வரும் வெளியீடு இலக்கு மொழிக்கு மாற்றப்படவேண்டும். இங்குச் சிக்கலான செயல்பாடு என்னவென்றால் உருபனியல் மற்றும் தொடரியல் பகுப்பு வெளியீடுகள் பெரும்பாலும் ஒன்றிற்கு மேற்பட்ட அர்த்தத்தை வெளிப்படுத்தி நிற்கும் என்பதாகும். அதாவது மூல மொழியில் இருக்கும் ஒரு வாக்கியத்திற்கு ஒன்றிற்கு மேற்பட்ட உருபனியல் பகுப்புகளோ தொடரியல் பகுப்புகளோ வருவது சாத்தியமாகும்; அதாவது ஒன்றிற்கும் மேற்பட்ட பொருண்மையியல் உருப்படுத்தங்கள் சாத்தியமாகும். ஒரு மூல மொழி வாக்கியத்திலிருந்து கிடைக்கும் ஒன்றிற்கு மேற்பட்ட தொடரியல்

பகுப்புகளில் ஒன்று தான் சரியானது. அந்தச் சரியான ஒன்று தான் இலக்கு மொழிக்கு மாற்றப்பட வேண்டும். எனவே பொருள் மயக்கம் நீக்கப்பட வேண்டிய கட்டாயம் ஏற்படுகிறது. இதைச் சொல் மயக்க நீக்கம் (Word Sense Disambiguation (WSD)) என்பர். சொல் மயக்க நீக்கம் சூழல் அடிப்படையில் செய்யப்பட வேண்டியதாகும். இதற்காகப் பல புள்ளியியல் மாதிரிகள் இருக்கின்றன. இத்தகைய புள்ளியியல் மாதிரிகளைப் பயன்படுத்தி சொல்மயக்கம் நீக்கிச் சரியான இலக்கணப் பகுப்பைப் பெற்று இலக்கு மொழிக்கு மாற்றம் செய்தால் மொழிபெயர்ப்பு செம்மையானதாக இருக்கும்.

இயந்திரமொழிபெயர்ப்புக்கான முன் நடவடிக்கைகளாகப் பின்வருவனவற்றைப் பட்டியலிடலாம்: தரவுத்தொகுதி (corpus) உருவாக்குதல், டோக்கன் பகுப்பான் (tokenizer), உருபனியல் பகுப்பாய்வி (morphological analyzer), தொடரியல் பகுப்பாபான் (syntactic parser), தொடர்கூறு பகுப்பான் (chunker), உருபனியல் அல்லது சொல் உருவாக்கி (morphological/word generator), தொடரியல் அல்லது வாக்கிய உருவாக்கி (syntactic/sentence generator), சொற்பொருள் மயக்கநீக்கி (word sense disambiguator) என்பன உருவாக்குதல், இருமொழிய அல்லது பன்மொழிய அகராதி உருவாக்குதல்.

தரவுத்தொகுதி தயாரித்தல் மிக முக்கியமான செயல்பாடாக மாறியுள்ளது. தரவுத்தொகுத் தயாரிப்பதும் அத்தரவுத் தொகுதியை இயந்திரமொழிபெயர்ப்புச் செயல்பாடுகளுக்கு வேண்டி அடையாளப்படுத்துவதும் அதில் வரும் சிக்கல்களையும் தீர்வுகளையும் இன்றியமையாத செயல்பாடுகளாகும். இயற்கைமொழி ஆய்விற்கும் இயந்திர மொழிபெயர்ப்பு ஆய்வுக்கும் தரவுத்தொகுதி முக்கியமான பகுதியாக அமைகிறது. ஒரு காலகட்டத்தில் மொழியியலில் கள ஆய்வுகள் மூலம் தரவுகள் திரட்டப்பட்டு, ஆராயப்பட்டு விதிகள் உருவாக்கப்பட்டன. இவ்விதிகள் குறைந்த அளவிலான தரவின் அடிப்படையில் அமைந்து முழுவதுமாக மொழி அமைப்பை விளக்காமல் போய்விட்டதன் காரணமாக இக்கள ஆய்வு குறைகள் நிறைந்த்தாய்க் கருதப்படல்லாயிற்று. இதற்கு மாறாகச் சாம்ஸ்கி என்பவர் அறிதிறனை அடிப்படையாகக் கொண்டு மொழியியல் விதிகள் அமைய வேண்டும் என்று வலியுறுத்திக் கூறினார். இந்த இருவித அணுகுமுறையிலும் சிக்கல்கள் இருந்துவந்தன. மொழியை ஒட்டுமொத்தமாக விளக்க கள ஆய்விலிருந்து பெறும் தரவோ அறிதிறன் அடிப்படையில் அமையும் தீர்மானங்களோ போதாது என்பதன் அடிப்படையில் புள்ளியியலைச் சார்ந்து விதிகளை அல்லது தீர்மானங்களை எடுக்கும்

நோக்கு முக்கியத்துவம் பெற்றது. இதன் காரணமாக மொழியியல் ஆய்வானது களத் தரவு ஆய்விலிருந்து தரவுத்தொகுதி ஆய்வுக்கு நீட்சியுற்றது. புள்ளியியல் அடிப்படையிலான தீர்மானங்கள் நம்பத்தகுந்ததாக அமைந்தால் தரவுத்தொகுதி தயாரித்தல் மிக முக்கியமான செயல்பாடாக அமைந்தது. இயற்கைமொழி ஆய்வின் எந்தச் செயல்பாட்டிற்கும் தரவுத்தொகுதி அடிப்படையிலான புள்ளியியல் ஆய்வு முக்கியமாகக் கருதப்பெற்றது. தமிழும் இம்முயற்சிகளிலிருந்து பின்வாங்கவில்லை. தமிழுக்கான தரவுத்தொகுதிகள் பல தயாரிக்கப்பட்டு இன்று பயன்பாட்டில் உள்ளன.

வேறுபட்ட வகையிலான அகராதிகள் தயாரிப்பதற்கும் சொற்களஞ்சியம் தயாரிப்பதற்கும் இயந்திர மொழிபெயர்ப்புக்கும் தரவுத்தொகுதி இன்றியமையாததாகும். இயந்திர மொழிபெயர்ப்புக்கு மிக முக்கியமாகக் கருதப்படும் சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை தரவுத்தொகுதியின் புள்ளியியல் விவரங்களின் அடிப்படையிலானதாகும். பேச்சு உரை மாற்றத்திற்கும் உரை பேச்சு மாற்றத்திற்கும் பேச்சு தரவுத்தொகுதிகள் கைக்கொடுக்கும். இவ்வாறு தரவுத் தொகுதியின் முக்கியத்துவம் குறித்து கூறிக்கொண்டே செல்லலாம்.

அடையாளப்படுத்தப்படாத தரவுத்தொகுதியை வைத்துக்கொண்டு எந்த இயற்கை மொழி ஆய்வையும் செய்ய இயலாது. தரவுத்தொகுதி சொல்வகைப்பாட்டிற்காக அடையாளப்படுத்தப்பட்டால்தான் அத்தரவுத்தொகுதி பயனுள்ளதாக அமையும். தமிழைப் பொறுத்தவரையில் பல சொல் வகைப்பாடுகள் இருப்பதைப் பார்த்தோம். தமிழ்த் தரவுத்தொகுதியானது சொல்வகைப்பாடுகளுக்காக அடையாளப்படுத்தப்பட்டால்தான் அவற்றைப் பயன்படுத்த முடியும். பலவகை சொல்வகைப்பாட்டு குழுமங்கள் வழக்கத்திலிருந்து வருகின்றன. ஆங்கிலத்திற்கென்று உருவாக்கப்பட்ட சொல்வகைப்பாட்டு குழுமங்களைப் பயன்படுத்துவது, இந்திக்கென்று உருவாக்கப்பட்ட சொல்வகைப்பாட்டு குழுமத்தைப் பயன்படுத்துவது, தமிழுக்கென்று தனி சொல்வகைப்பாட்டு குழுமத்தை உருவாக்கிப் பயன்படுத்துவது போன்ற செயல்பாடுகள் இன்று நடைமுறையில் உள்ளன. இந்திய மொழிகளுக்காகச் சொல்வகைப்பாட்டுக் குழுமத்தை IIIT ஹைதராபாத் நிறுவனம் உருவாக்கிப் பயன்படுத்தி வருகின்றது. இந்திய மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பில் இந்நிறுவனத்தின் சொல்வகைப்பாட்டு அடையாளக் குழுமம் பயன்படுத்தப்பட்டு வருகிறது. இது குறைபாடு உள்ளதாகத் தோன்றுகிறது. இச்சொல்வகைப்பாட்டு அடையாளக் குழுமம் இந்தி

மொழியின் அடிப்படையில் உருவாக்கப்பட்டதால் தமிழ் போன்ற திராவிட மொழிகளுக்கு சரியாகப் பொருந்தவில்லை. மைசூரில் உள்ள இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்திய மொழிகளுக்கான குறிப்பாக இந்தி மொழிக்கான சொல்வகை அடையாளக் குழுமத்தை உருவாக்கிப் பயன்படுத்தி வந்தது. சென்னையிலுள்ள AUKBC நிறுவனம் தமிழுக்கென்று தனியாகச் சொல்வகைப்பாட்டு அடையாளக் குழுமத்தை உருவாக்கிப் பயன்படுத்தி வந்தது. இந்திய மொழிகளுக்கான சொல்வகை அடையாளக் குழுமத்தை Eagles குறிப்பிடும் பரிந்துரைகளின் அடிப்படையில் தரப்படுத்தும் முயற்சி பெங்களூரில் உள்ள MSRI நிறுவனத்தால் மேற்கொள்ளப்பட்டு சொல்வகைப்பாடு அடையாளக்குழு நிர்ணயிக்கப்பட்டுள்ளது. இந்திய அரசின் நிதி நல்கையில் நடைபெற்றுவரும் இயந்திரமொழிபெயர்ப்புத் திட்டங்களுக்காக இத்திட்டங்களின் கூட்டுக்குழுகள் (consortia) ஒன்றுகூடி விவாதித்து இந்திய இயந்திர மொழிபெயர்ப்புத் திட்டங்களுக்கான ஒரு பொதுவான சொல்வகைப்பாட்டு குழுமத்தையும் (BIS tagset), தொடர்கூறு வகைப்பட்டுக் குழுமத்தையும் நிர்ணயித்து வெளியிட்டுள்ளனர்.

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி தொடர்கூறுகளுக்காக அடையாளப்படுத்தப்பட வேண்டும். தொடர்கூறுகளைக் கண்டறிந்த பின்னர்தான் நாம் வாக்கியதில் இத்தொடர் கூறுகளின் செயல்பாட்டுகளை உணர்ந்து மூல மொழியை இலக்கு மொழிக்கு மாற்ற இயலும். இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்புத் திட்டந்தில் இதற்கென்று தொடர் களின் குழுமம்/கணம் நிறுவப்பெற்று புள்ளியியல் அடிப்படையில் தரவுத்தொகுதியை அடையாளப்படுத்தும் செயல்பாடும் நிர்ணயிக்கப்பட்டுள்ளது. ஆழமில்லாப் பகுத்துக் குறித்தலின் ஒரு பகுதியாக அமையும் தொடர்கூறு பகுத்தல் முழுத் தொடரியல் பகுப்பாய்வைக் காட்டிலும் சிறந்ததாகக் கருதப்படுகின்றது.

பொதுவான இயந்திரங்களை (உருபனியல் பகுப்பாய்வு இயந்திரம், சொல்வகை அடையாளப் படுத்தும் இயந்திரம், தொடர்கூறு பகுக்கும் இயந்திரம், சொல் உருவாக்கி இயந்திரம், வாக்கிய உருவாக்கி இயந்திரம்) போன்ற இயந்திரங்களை எல்லா மொழிகளுக்கும் பயன்படுமாறு உருவாக்கி குறிப்பிட்ட மொழிக்கென்று மாறுதல்கள் செய்து பயன்படுத்துவது செலவுச் சிக்கனத்திற்கும் காலச் சிக்கனத்திற்கும் வழிவகுக்கும். தற்போது உள்ள சொல்வகைப்பாட்டுக் குழுமம் புதிய தரவுகளின் அடிப்படையில் மாற்றம் செய்யப்படவேண்டும்.

1.6. கவனத்தில் கொள்ளவேண்டிய செய்திகள்

பின்வரும் செய்திகளைகருத்தில் கொள்ளவேண்டும்:

- முழுமையான பகுத்துக்குறித்தலைக் காட்டிலும் ஆழமில்லாப் பகுத்துக் குறித்தல் இயந்திர மொழிபெயப்புக்குப் போதுமானது.
- பொதுவான பொருண்மைக்கள உரைகளைக் காட்டிலும் ஒரு குறிப்பிட்ட பொருண்மைக் களத்தைச் சார்ந்த உரைகளை (சுற்றுலா, உடல் நலம் போன்ற பொருண்மைக்களங்கள்) இயந்திர மொழிபெயர்ப்பு செய்வது எளிதாகும்.
- இந்திய மொழிகளுக்கு என்று ஒரு பொதுவான சொல்வகைப்பாட்டுக் குழுமத்தை உருவாக்கிச்செயல்படவியலும்.
- சொல்வகைப்பட்டுக் குழுமங்கள் தேவைக்குத் தகுந்தவாறு குறைவாக இருந்தால் போதுமானதாகும்.
- சொல்வகைப்பாட்டின் போது ஏற்படும் பொருண்மை மயக்கங்கள் தொடர்கூறு பாகுபடுத்தலின் போது நீக்கப்படும்.

1.7. கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழி மாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்ததைய கணிப்பொறி வழி மொழி பெயர்ப்புப் பணிகளில் எல்லைகள் இரண்டாக வரையறுத்து உள்ளன.

1.மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் அமைப்புகளின் தொகுப்பாகும். ஒரு வட்டாரத்தில் பயன்படுத்தக்கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்தக்கூடிய உறுதியில்லை. தமிழ் போன்ற மொழிகளில் மொழி நடை இலக்கியமொழி, பேச்சுமொழி என இருவகையாகப் பிரிக்கப்படுகிறது. பேச்சுமொழியைக் கணிப் பொறியாக்கம் செய்வது என்பது சுலபமான பணியன்று. எனவே, மொழிப்பெயர்ப்புப் பணியில் இலக்கிய மொழிகள் மட்டுமே எடுத்துக்கொள்ளப்படுகின்றன.

2.செய்யுள், கவிதை போன்றவைகளில் பொருளுக்குத் தரும் முக்கியதுவத்தைவிட கருத்தைப் புலப்படுத்தும் உத்தி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகின்றன. இவைகள் மனித உளவியலைப் பிரதிப்பலிப்பன. இவ்வுளவியல் கூறுகளைக் கணிப்பொறி வழி வெளிப்படுத்துவது என்பது இயலாத ஒன்றாகும். இதேபோல் இலக்கியங்களிலும் பொருளைவிட உளவியல் கூறுகளுக்கு தான் முக்கியத்துவம் தரப்படுகின்றது. எனவே, செய்யுள்கள், கவிதைகள்,

பாடல்கள் மற்றும் இலக்கியச் செய்திகள் ஆகியவைகளைக் கணிப்பொறி வழி மொழிபெயர்ப்பதற்குத் தொடக்கநிலையில் பயிற்சி எடுத்துக்கொள்ளப்படவில்லை. அறிவியல் தொழில்நுட்ப சார்ந்த படைப்புகளில் பொருள்களுக்கே முக்கியத்துவம் தரப்படுகிறது. இங்கு கோபம், வீரம், தாபம், மகிழ்ச்சி போன்ற உணர்வுகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணிகளுக்குப் பொதுவாக அறிவியல் மற்றும் தொழில்நுட்பத்தைச் சார்ந்த படைப்புகளே எடுத்துக்கொள்ளப்படுகின்றன.

இதேப் போன்று மொழிகளின் தன்மையைப் பொறுத்துக் கணிப்பொறி மொழிபெயர்ப்பினை இருபெரும் பிரிவுகளாகப் பிரிக்கலாம்:

1.இருதொடர்புடைய மொழிகளுக்கிடையே அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Interlingual Method of Machine Translation): இலக்கண அமைப்பிலும் பொருள் அமைப்பிலும் தொடர்பும் நெருக்கமும் உடைய இருமொழிகளுக்கிடையே செய்யப்படும் கணிப்பொறி மொழிப்பெயர்ப்பு இவ்வகையைச் சார்ந்தாகும்.

2.இடைநிலை மொழி அடிப்படையில் அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Intermediate Language Based Machine Translation) இலக்கணம் மற்றும் மொழியமைப்பில் முற்றிலும் வேறுபட்ட இருமொழிகளுக்கிடையே கணிப்பொறி வழி மொழிபெயர்ப்பை நடைமுறைப்படுத்த இம்முறை உதவுகிறது. இம்முறையில் மூலமொழியில் உள்ள சொற்றொடர்கள் இடைநிலை மொழி வழி இலக்கு மொழிக்கு மாற்றம் செய்யப்படுகின்றது.

1.8. கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழிபெயர்ப்புப் பணியைக் கணிப்பொறி வழி மேற்கொள்கின்ற பொழுது கணிப்பொறியின் சிறப்புத் தன்மை, துல்லியம், வேகம், நினைவகக் கொள்திறன், தகவல் தளம் உருவாக்கும் வசதிகள் போன்றவை மொழிபெயர்ப்புத் திறனை நிர்ணயம் செய்யும். ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழிபெயர்ப்பு செய்யும் போது பன்மொழித் திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வழியமைக்கப்படுகிறது. மனித மொழிபெயர்ப்புப் பணிக்குக் கூட ஒருவருக்கு இருமொழியிலும் திறமை இருக்கவேண்டும் என்ற அவசியம் இல்லை.

ஒரு மொழியின் வளம், தன்மை ஆகியவற்றைக் கணிப்பொறிக்கு ஏற்றாற்போல் கணிதப்படுத்துவது என்பது எளிமையான பணி அல்ல. எனவே, கணிப்பொறி வழி மொழி

பெயர்ப்புத் திட்டங்கள் சுமார் 50 ஆண்டுகளுக்கு முன்னர் தொடங்கப்பட்டிருப்பினும் இன்னும் முழுமைபெறவில்லை. மேலும் கணிப்பொறி அறிவியலில் தொழில் நுட்ப வளர்ச்சி கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிக்கு மகிவும் பயனுள்ளதாக அமைகின்றது. இத்தகைய பணிகளின் நிறை குறைகளைக் கீழேக் காணலாம். மொழிக்கு இடைடே உள்ள சொல் வளம் இலக்கணம் கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவைகளில் மொழிக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிபெயர்ப்புப் பணியின் சிக்கல்கள் குறையும்.

உருபனியல் அமைப்பை பொறுத்தவரையில் மொழிகளின் பண்புகள் இருவேறு திசைகளில் மாறுகின்றன:

அ. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதைப் பொறுத்து.

ஆ. ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இருந்தாலும் அதற்கு மேலே இருந்தாலும் அவைகளைத் துல்லியமாகப் பிரிக்க இயலும் என்பதைப் பொறுத்து.

ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இந்தால் அந்த மொழியைத் தனிநிலை உருபன்கட்டு மொழி (isolating Language) மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை பன்நிலை உருபன் கட்டுமொழி (Poly Synthetic) என்கிறோம். சிபேரியன் மொழி பன்நிலை உருபன் கட்டுமொழிக்குச் சிறந்த உதாரணமாகும். வியட்நாமிஸ் மற்றும் கேண்டுனஸ் ஆகிய மொழிகள் தனிநிலை உருபன்கட்டு (isolating) மொழிக்குச் சிறந்த உதாணரங்களகும். எனவே இத்தகைய மொழிக்குச் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் வகுக்கின்ற பொழுது மொழியின் உருபனியல் பற்றிய விரிவான தகவல் அவசியம் ஆகும்.

மொழியின் தொடரியல் அமைப்புகள் (Syntactic Structure) மூன்று வகைகளாய் பிரிக்கப்படுகின்றன.

1. SVO அமைப்பு

ஜெர்மன், பிரஞ்சு, ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

தமிழ், இந்தி, ஜப்பான் மொழிகள் SOV அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரீஸ், அரேபி, ஷிபிரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளன.

SVO அமைப்பில் உள்ள மொழிகளில் முன்னுருப்புகள் (Preposition) பயன்படுத்தப்படுகிறது. எனவே கணிப்பொறி வழி மொழிபெயர்ப்புத் திட்டங்களுக்கு மொழிகளுக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள், உறவுகள் போன்றவை ஆராயப்பட வேண்டும்.

இராமனின் வீடு என்ற தொடரில் *வீடு* என்பது தலைச்சொல் (Head Word) ஆகும். *இராமன்* என்பது சார்புச் சொல் (Dependent Word) ஆகும். தலைச்சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையில் உள்ள உறவை *-இன்* என்ற ஒட்டு விளக்குகிறது. இது போன்று தலைச் சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையிலான முறையைப் பொறுத்து மொழி தலைக்குறித்தல் மொழி, சார்புக்குறிதல் மொழி எனப் பிரிக்கப்படும். சார்பு குறித்தல் மொழியில் ஒட்டு சார்புச் சொல்லுடனும் தலைக்குறித்தல் மொழியில் ஒட்டு தலைச் செல்லுடனும் தரப்படுகின்றன. ஆங்கில மொழி தலைக்குறித்தல் மொழியைச் சார்ந்ததாகும். ஹங்கேரிய மொழி சார்புக்குறித்தல் மொழியைச் சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு விளக்கும்.

English	:	The Man's house
Hungarion	:	The man house his
House	:	Head word (தலைச்செல்)
The Man	:	Dependent word (சார்புச்சொல்)
-s	:	Affix (ஒட்டு)

எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியை மேற்கொள்கின்ற போது தலைச்சொல்லுக்கும், சார்பு சொல்லுக்கும் இடையே உறவு முறைகள் எவ்வாறு குறிக்கப் பெறுகின்றன என்பது பற்றியெல்லாம் ஆராயவேண்டும்.

மூலமொழியில் உள்ள ஒரு சொல்லின் வகைப்பாடும் இலக்கு மொழியிலுள்ள அதற்கு நிகரான சொல்லின் வகைப்பாடும் ஒன்றாக இருக்கவேண்டியத் தேவையில்லை. எடுத்துக்காட்டாக ஒரு மொழியிலுள்ள வினைச்சொல் பிறமொழிக்கு மொழி பெயர்ப்பு செய்யப்படும் போது மொழிப்பெயர்ப்புச் சொல் வினைச்சொல்லாக இருக்கவேண்டிய தேவையில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்புப் பணியில் உள்ள சிக்கல்களாகும்.

ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கானச் சொல் விளக்கமானதாகவோ கருக்கமானதாகவோ இருக்கலாம். எடுத்துக்கட்டாக *சகோதரன்* என்ற உறவைக் குறித்து ஆங்கிலத்தில் *brother* என்ற சொல் உள்ளது. மூத்த சகோதரனை ஆங்கிலத்தில் *elder brother* என்று சொல்லுகிறோம். இவ்வாறு வயது மூத்தவர் இளையவர் என்பதைக் குறிக்க *elder*, *younger*

என்ற அடைச் சொற்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் *அண்ணன்* என்றும் *தம்பி* என்றும் தனித்தனிச்சொற்கள் உள்ளன. *சாகோதரன்* என்ற உறவை மேலும் வகைப்படுத்த முதியவர், இளையவர் என்ற நிலையில் *அண்ணன்*, *தம்பி* என்ற இருச்சொற்கள் உள்ளன. ஆனால் ஆங்கிலத்தில் இவ்வகையிலான வேறுபாடு என்பது இல்லாதால் உறவை வேறுப்படுத்த *elder*, *younger* என்ற சொற்களைப் பயன்படுத்த நேர்கிறது.

மூலமொழியில் உள்ள செய்திகளில் கலாசாரப் பின்னணி மொழிப்பெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்கு மொழியின் கலாச்சாரப் பண்புகளுக்கு ஏற்ப மாற்றித்தரத்தக்க அளவு இருத்தல்வேண்டும். இருமொழிகளுக்கு இடையே உள்ள கலாச்சார தொடர்புகள் மொழிபெயர்ப்பு செய்ய இயலுமா? இயலாதா? என்பதைக் தெளிவுபடுத்த உதவும். மேலும் சபீர்-ஊர்ஃப் கொள்கையின் படி ஒருவர் பேசும் மொழி சிந்திக்கும் திறனைக் கூட்டுவதாகவோ குறைப்பதாகவோ உள்ளது. எனவே மொழிபெயர்ப்பில் இத்தகைய உளவியல் அடிப்படையிலானச் சிக்கல்கள் உள்ளன.

தொடரியல் பகுப்பாய்வி பல பணிகளுக்காக உருவாக்கப்படுகின்றன. ஆனால் இயந்திர மொழிபெயர்ப்பிற்காகப் பயன்படுத்தப்படும் பகுப்பாய்வு மற்ற பகுப்பாய்வை விட சிறிது வேறுபடுகிறது. இங்குத் தொடரின் அல்லது வாக்கியத்தின் பொருளை அடிப்படையாகக் கொண்டு பகுப்பாய்வு செய்யப்படுகிறது. எனவே இயந்திர மொழிபெயர்ப்பிற்கு உருவாக்கப்படும் பகுப்பாய்வு தொடரியல் அமைப்பை மட்டும் சாராது பொருளைக் சார்ந்ததாகவும் அமைகிறது. இத்தகைய பகுப்பாய்வுகளை உருவாக்க அறிவு அடிப்படை (Knowledge Base) தேவைப்படுகின்றது.

இயந்திரம் வழி மொழிபெயர்ப்பு செய்கின்ற பொழுது ஏதேனும் சிக்கல்கள் இருந்தாலோ அல்லது கணிப்பொறிக்குக் கூடுதல் கதவல் தேவைப்பட்டாலோ மனித உதவியை எதிர்நோக்குகின்றது. இத்தகைய சூழல்களில் சிக்கலைப் புரிந்துக்கொண்டு அதற்கேற்றத் தீர்வு உரிய முறையில் வழங்கப்பட்டவேண்டும். எனவே இயந்திர மொழிப்பெயர்ப்புத் திட்டத்தை இயந்திர மனித உரையாடலுக்கு ஏற்றாற்போல் வடிவமைக்கவேண்டியது அவசியம் ஆகும். பெரும்பாலும் படைப்பாற்றல் திறனைச் செய்திகளில் வெளிப்ப்படுத்தும் பொழுது கணிப்பொறிக்குக் கூடுதல் செய்திகள் தேவைப்படுகிறன்றன.

பெரும்பாலான கணிப்பொறி வழி மொழிபெயர்ப்புகளிலும் உள்ளீடு செய்யும் செய்திகளை முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும் இலக்கு மொழியின் தன்மையும்

கருத்தில் கொள்ளப்படுகின்றன. மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழிபெயர்ப்பது, பின்னர் மொழி பெயர்ப்பைப் படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகின்றதா என்பதை ஆராய்வது, அப்படி இல்லையெனில் மூலமொழி செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிபெயர்ப்பு செய்வது, சரியான கருத்து வெளிப்படுவதுவரை இத்தகைய ஆய்வை மீண்டும் மேற்கொண்டு மொழி பெயர்ப்பிற்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் அளித்து தேவையான திருத்தங்களைச் செம்மையாக மேற்கொள்வது. இம்முறை கணிப்பொறியின் நேரத்தை மிச்சப்படுத்துகிறது; மொழிபெயர்ப்பின் வேகத்தை அதிகப்படுத்துகிறது. மேலும் ஒரு மொழியிலிருந்து பல மொழிகளுக்கு மொழிபெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப்பணித் திறமையானதாக அமைகிறது.

மொழிபெயர்ப்புப் பணிகளுக்காக உருவாக்கப்படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லாது பல கூடுதல் செய்திகளைக் கொண்டதாக இருத்தல் வேண்டும். பிறவகை அகராதிகளை விட துறை சார்ந்த அகராதிகள் மொழிபெயர்ப்புப் பணியை எளிமைப்படுத்துகிறது. மூலமொழியில் உள்ள செய்திகளைப் பொறுத்து அது தொடர்புடைய அகராதியைப் பயன்படுத்திக்கொள்ளலாம்.

மொழிபெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத் தொடர்கள் மேலும் அடிக்கடி பயன்படுத்தக்கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்படவேண்டும்.

கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணி என்பது ஒரே மூச்சில் ஒட்டு மொத்தமாகச் செய்யக்கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கப்படவேண்டும். அதன் செயல் திறனை அறியவேண்டும். பணியின் தன்மையை மதிப்பிடவேண்டும். இதற்கு ஏற்றார் போல் தேவைப்படின் மென்பொருளைக் கூடுதல் வசதிகளுடன் மேம்படுத்தப்பட வேண்டும். இம்முறைகளை மனிதனுக்கு நிறைவழிக்கும் கொள்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்பச் செய்ய வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்புக் கடுமையானப் பணியாகும். இதற்கு மனிதமொழி பெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (Model) வழிமுறைகளும் அவசியமாகும். கணிப்பொறி வழி மொழிபெயர்ப்பு எண்ணற்ற முயற்சிகளால் உருவாக்கப்படுகின்ற ஆக்கப்பூர்வமான பணியாகும். இதற்கு அறிவியல் திறன் மட்டுமல்லாது மொழிபெயர்ப்பிற்குத் தேவையான கலைத்திறனும் அவசியமாகும்.

தற்பொழுது உலகளவில் நடைபெற்றுள்ள மொழிபெயர்ப்புப் பணிகள் மனதுக்கு நிறைவழிக்கக்கூடிய வகையில் உள்ளன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்கு உதவுகின்றன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்குப் போதுமானதாகும்.

மனிதனின் படைப்பாற்றல் கற்பனைத்திறன், உளவியல்பாங்கு போன்ற செய்திகளால் பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியில் இத்தகையக் கூறுகளைச் சேர்க்கவேண்டி உள்ளது. அதற்கு மனித நுண்ணிறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன.

1.9. சுருக்கவுரை

இங்கு கணிப்பொறி வழி இயந்திர மொழிபெயர்ப்பின் தேவை, இயந்திரமொழிபெயர்ப்பின் வகைகளும் அணுகுமுறைகளும், இயந்திரமொழிபெயர்ப்புக்கான இலக்கண வடிவமைப்புகள், இயந்திர மொழிபெயர்ப்புக்கான முன்னேற்பாடுகள், கவவனத்தில் கொள்ளவேண்டிய செய்திகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும் என்ற தலைப்புகளில் விளக்கங்கள் தரப்பட்டுள்ளன.

இயல் 2

இயந்திர மொழிபெயர்ப்பின் மைல்கற்கள்

2.1. அறிமுகம்

ஹட்சின்ஸ் தன்னுடைய பல கட்டுரைகளிலும் நூல்களிலும் இயந்திரமொழிபெயர்ப்பின் வளர்ச்சி குறித்து விரிவாகப் பேசுகின்றார் (Hutchins 1986-2010, Hutchins and Somers 1992). இயந்திர மொழிபெயர்ப்பு அமைப்பொழுங்கின் வளர்ச்சி கடந்த நூற்றாண்டுகளில் செய்யப்பட்ட ஆராய்ச்சி முன்னேற்றம் குறித்தக் காரணிகளால் ஊக்கப்படுத்தப்பட்டுள்ளது. இயந்திர மொழிபெயர்ப்பின் உண்மையான பிறப்பு இரண்டாவது உலகப்போருக்குப்பின் டிஜிட்டல் கணிப்பொறியின் வரவால் விளைந்தது. ஹட்சின்ஸ் (Hutchins 2010) இயந்திர மொழிபெயாப்பின் வளர்ச்சியை சில கால கட்டங்களாகப் பகுத்து ஆய்ந்துள்ளார். பின்வருவன சில முக்கியமான காலகட்டங்களாகும்.

1. முதல் கட்டம் (1933-1956)
2. இரண்டாம் காலகட்டம் (1956-1966)
3. மூன்றாம் காலகட்டம் (1967-1976)
4. நான்காம் காலகட்டம் (1976-1989)
5. ஐந்தாம் காலகட்டம் (1989-இலிருந்து)

2.2. முதல் கால கட்டம் (1933லிருந்து 1956 வரை)

ஹட்சின்ஸ் இக்காலகட்டத்தை 'முந்திய நிலை மற்றும் முன்னோடிகள்' (Precusors and pioneers) என்று குறிப்பிடுகின்றார். நாம் இயந்திர மொழிபெயர்பை 17ஆம் நூற்றாண்டின் உலகப்பொதுமொழிகள் மற்றும் 'இயந்திரம்சார்' அகராதிகளுக்குக் கொண்டு செல்லலாம் என்றாலும் இருபதாம் நூற்றாண்டுவரை இது நடைமுறைப் படுத்தப்படவில்லை. 1933-இல் தான் இரண்டு காப்புரிமைகள் பிரான்சிலும் ரஷ்யாவிலும் முறையே ஜார்ஜெஸ் ஆஸ்ட்ரோனிக்கும் (Georges Arstrouni) பெட்ர் ட்ரொஜன்ஸ்கிஜ்-க்கும் (Petr Trojanskij) வழங்கப்பட்டன. ஆஸ்ட்ரோனியின் காப்புரிமை இயந்திரம்சார் பன்மொழிய அகராதியாகச் செயல்பட இயலும் பொதுநோக்க இயந்திரத்திற்கு ஆகும். ட்ரொஜன்ஸ்கிஜியின் காப்புரிமையும் அடிப்படையில் இயந்திரம்சார் அகராதிக்கு ஆகும்; இது பன்மொழிய மொழிபெயர்ப்புக் கருவியில் 'உலகப்பொது' (எஸ்பெராந்தோ அடிப்படை (Esperanto-based)) குறியீடுகளைப் பயன்படுத்தி குறியாக்கம்

செய்வதற்கும் இலக்கணச்செயல்பாடுகளைப் பொருள்கோள் செய்வதற்கும் முன்மொழிவுடன் மேம்படுத்தப்பட்டுள்ளது.

இந்த முன்னோடிகள் பற்றி ஆண்ட்ரூ பூத்திற்கும் (Andrew Booth) வாரன் வீவருக்கும் (Warren Weaver) அவர்கள் 1946இலும் 1947இலும் சந்தித்தபோது தெரிந்திருக்கவில்லை. அவர்கள் இயற்கை மொழிகளை மொழிபெயர்ப்பு செய்வதற்குப் புதிதாகக் கண்டுபிடிக்கப்பட்ட கணிப்பொறிகளைப் பயன்படுத்தும் முதல் தற்காலிகக் கருத்தை முன்வைத்தனர். பூத் 1948இல் ரிச்சர் எச். ரிச்சென்ஸுடன் (Richard H. Richens) (Commonwealth Bureau of Plant Breeding and Genetics, Cambridge,UK) ஒரு இயந்திரம் சார் அகராதிக்கு வேண்டி உருபனியல் பகுப்பாய்வில் ஈடுபட்டார். இச்சமயத்தில் இயந்திரம்சார் மொழிபெயர்ப்பு (1960 தொடக்கம் வரை அறியப்பட்டது) பல எண்ணிக்கையிலான மக்களுக்குத் தன்னிச்சையாகத் தோன்றியது. ஜூலை 1949இல் வாரன் வீவர் (இயக்குநர், ராக்ஃபெல்லர் ஃபவுண்டேஷன் (Rockefeller Foundation)) தமது மறைகுறியீட்டியல் (cryptography) அறிவு, புள்ளியியல் (statistics), தகவல் கோட்பாடு (information theory), தர்க்கவியல் (logic) மற்றும் மொழி உலக பொதுமைகள் (language universals) அடிப்படையில் பொருண்மை மயக்கத்தின் (ambiguity) (அல்லது 'பல் பொருண்மைகள்') சிக்கல்களைக் கையாளுவதற்குக் குறிப்பிட்ட முன்மொழிவுகளை முன்வைத்தார்.

பின்னர் மே 1951இல் எஹொஷுவா பர்ஹிலெல் (Yahoshua Bar-Hillel) மாசசூசெட்ஸ் தொழில்நுட்ப நிறுவனத்தில் (Massachusetts Institute of Technology (MIT)) ஆய்வுக்கு வேண்டி பணியில் அமர்த்தப்பட்டார். இந்த விஷயத்தில் விருப்பம் உள்ள எல்லோரையும் சந்தித்த பின்னர் அவர் தற்போதைய நிலைமை பற்றி ஒரு அறிக்கை எழுதினார்; அதில் அவர் இயந்திர மொழிபெயர்ப்பு பற்றிய கேள்விகளுக்குச் சில அடிப்படை அணுகுமுறைகளைச் சுருக்கமாகக் கூறினார்; ஜூன் 1952இல் அவர் (எம்ஐடி-இல்) முதல் இயந்திர மொழிபெயர்ப்பு கருத்தரங்கைக் கூட்டினார்; இக்கருந்தரங்கில் இந்தப் புலத்தில் ஏற்கனவே செயல்திறனுடைய எல்லோரும் கலந்துகொண்டனர். நல்ல தரமான மொழிபெயர்ப்பின் முழு தானியக்கம் கிட்டத்தட்ட சாத்தியமற்றது, கணினி செயல்பாட்டிற்கு முன்னரோ பின்னரோ மனிதத் தலையீடு (முன்-திருத்தியமைத்தலும் பின்-திருத்தியமைத்தலும் (pre-editing and post-editing)) தேவை என்பது முன்னரே தெளிவானதாகும். சிலர் இது ஒரு இடைக்கால நடவடிக்கை என்றும் பெரும்பாலானோர் இது எப்போதும் தேவைப்படும் என்றும் நம்புகின்றனர். கருத்தரங்கில் முன் திருத்தியமைத்தலுக்கும் பின் திருத்தியமைத்தலுக்கும் பொருண்மை மயக்கச் சிக்கல்களைக்

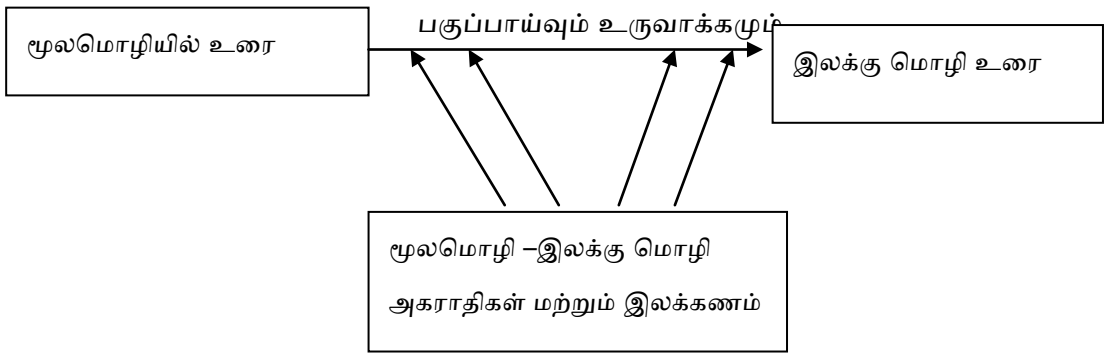
குறைக்க நுண் பொருள்விளக்கச் சொற்கோவைகளும் (micro glossaries) (பொருத்தமான இலக்கு மொழிச் சொற்களைத் தெரிந்தெடுக்க) மற்றும் சில வகைத் தொடரியல் அமைப்பு ஆய்வுக்கும் பல கருத்துக்கள் முன்வைக்கப்பட்டன. ஜார்ஜ்டவுணிலிருந்து (Georgetown University) வந்திருந்த லியோன் டொஸ்டெர்ட் (Léon Dostert) ஆய்வு நிதிநல்கையைக் கவர வேண்டி இயந்திர மொழிபெயர்ப்பின் சாத்தியம் பற்றிய பொது நிரூபணம் தேவை என்று வாதாடினார்.

இதன்படி அவர் 1954 ஜனவரி 7இல் ஒரு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையின் முதல் நிரூபணத்தில் விளைந்த ஒரு ஆய்வுத்திட்டத்தில் ஐபிஎம்-உடன் ஒத்துழைத்தார். இது ஐபிஎம்-இன் பீட்டர் ஷெரிடனில் (Peter Sheridan) மற்றும் ஜார்ஜ்டவுணின் பால் கார்வின் (Paul Garvin) ஆகியோரின் கூட்டுமுயற்சியாகும். கவனத்தோடு தெரிந்தெடுக்கப்பட்ட 49 ருஷ்ய மொழி வாக்கியங்கள், மிக எல்லைப்படுத்தப்பட்ட 250 சொற்றொகையையும் 6 இலக்கண விதிகளையும் பயன்படுத்தி ஆங்கிலத்தில் மொழிபெயர்க்கப்பட்டன. இந்த நிரூபணம் ஐக்கிய அமெரிக்காவில் தகவல் ஊடகத்தின் கூடுதல் கவனத்தை ஈர்த்தது. இந்த ஒழுங்குமுறைக்கு அறிவியல் மதிப்பு இல்லாதிருந்தாலும் அதன் வெளியீடு ஐக்கிய அமெரிக்காவில் இயந்திர மொழிபெயர்ப்புக்கு அதிக அளவில் நிதிநல்கையைத் ஊக்குவிக்கும் அளவுக்கு ஈர்க்கக்கூடியதாக இருந்தது; மேலும் உலக முழுவதும் பிற இடங்களிலும் குறிப்பாக ஐக்கிய ருஷ்யாவிலும் இயந்திர மொழிபெயர்ப்பு திட்டங்களைத் தொடங்குவதற்கு ஊக்கமளிப்பதாய் அமைந்தது.

அதே ஆண்டில், வில்லியம் லாகே (William Locke) மற்றும் 1953இல் எம்ஐடி-இல் பார்-ஹில்லெல்-ஐ (Bar-Hillel) பின்தொடர்ந்த விக்டர் யங்வே (Victor Yngve) என்போரால் முதல் இயந்திரம்சார் மொழிபெயர்ப்பு (Mechanical Translaton) என்ற ஆய்விதழ் நிறுவப்பட்டது. இது 1970இல் அதன் முடிவுக்கு முன்னர் வரை சில முக்கியமான ஆய்வுக்கட்டுரைகளை வெளியிட்டது. இந்த ஆண்டில் தான் இயந்திர மொழிபெயர்ப்பில் முதல் முனைவர் பட்ட ஆய்வேடு அந்தோணி ஜி. ஓட்டிங்கர்-இன் (Anthony G. Oettinger) ருஷ்யமொழி இயந்திரம்சார் அகராதி (Russian mechanical dictionary) பற்றிய ஆய்வு சமர்ப்பிக்கப்பட்டது. 1954 மற்றும் 1955 ஆண்டுகளில் இங்கிலாந்தில் கேம்பிரிட்ஜில் மார்க்ரெட் மாஸ்டர்மான் (Margaret Masterman) கீழ் ஒரு குழுவும் மிலானில் (Milan) ஸில்வியோ செகாட்டோவின் (Silvio Ceccato) கீழ் ஒரு குழுவும் துல்லியமான இயக்கவியல் மற்றும் கணினி தொழில்நுட்ப நிறுவனத்திலும் (Institute of Precise Mechanics and Computer Technology) லெனின்கார்ட் பல்கலைகழகத்தில் (Leningrad University) பயன்பாட்டுக் கணக்கியல் நிறுவனத்திலும் (Institute of Applied Mathematics) முதல் ருஷியமொழி குழுக்களும்

நிறுவப்பட்டன; மேலும் சீன மொழி, ஜப்பான் மொழி இவற்றில் பல்வேறு திட்டங்கள் தொடங்கப்பட்டன. 1955இல் முதல் இயந்திர மொழிபெயர்ப்பு நூல் தோன்றியது; இது வீவரின் (1949) மெம்மாரண்டத்தையும் (memorandum) 1952இல் கருத்தரங்களில் தரப்பட்ட சில கட்டுரைகளையும் பார்-ஹில்லெல் (Bar-Hillel), டோஸ்டெர்ட் (Dostert), (Oettinger) ரெய்ஃப்ளெர் (Reifler) மற்றும் யங்வே (Yngve) என்போரின் பிற பங்களிப்புகளையும் உட்படுத்தி லோகெ மற்றும் பூத் (Locke and Booth 1955) என்பவர்களால் பதிப்பிக்கப்பட்ட தொகுதி ஆகும்.

கீழே தரப்பட்டுள்ள படம் இக்காலகட்டத்தில் உருவாக்கப்பட்ட இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் பொதுவாக மூலமொழிச் சொற்களுக்கு இணையான சொற்களை இருமொழி அகராதி பார்த்துக் கண்டுபிடித்து மூல மொழி போலவே வரிசைப்படுத்தித் தருவதைச் செயல்முறையாகக் கொண்டிருந்தது மற்றும் தொடரியல் பகுப்பானைப் பயன்படுத்தவில்லை என்பதைத் தெளிவுபடுத்தும்.



2.3. இரண்டாம் கால கட்டம் (1956லிருந்து 1966 வரை)

ஹட்சின்ஸ் (Hutchins 2015) இக்காலகட்டத்தை "அதிக எதிர்பார்ப்புகளும் ஏமாற்றங்களும்" (High expectations and disillusion) என்று குறிப்பிடுகின்றார். இயந்திரம்சார் மொழிபெயர்ப்பு ஆராய்ச்சி தொடங்கியபோது, தற்போதைய மொழியியலில் இருந்து சிறிய உதவி கிடைத்தது. இதன் விளைவாக, 1950கள் மற்றும் 1960களில், ஆராய்ச்சி நெறிமுறைகள் இரண்டு எதிரெதிரான அணுகுமுறைகளாகப் பிரிய முனைந்தன; ஒருபுறம் அனுபவ சோதனை மற்றும் பிழை அணுகுமுறைகள்; இவை பெரும்பாலும் இலக்கண மற்றும் அகராதி சீர்மைகளின் 'கண்டுபிடிப்பு'க்காக (discovery) புள்ளியியல்சார் முறைகளைப் (statistical methods) பின்பற்றின; இது கணினி அடிப்படையில் பயன்படுத்தக்கூடியது ஆகும். மறுபுறம் அடிப்படை மொழியியல்

ஆராய்ச்சியில் திட்டங்களை உள்ளடக்கிய கோட்பாட்டு அணுகுமுறைகள்; இவை உண்மையில் பிற்காலத்தில் 'கணினி மொழியியல்' என்று அழைக்கப்பட்ட ஆராய்ச்சியின் ஆரம்பம் ஆகும். முரண்பாட்டு முறைகள் (வேற்றுநிலை முறைகள்) அந்த நேரத்தில் வழக்கமாக முறையே 'முரட்டுத்தனம்' (brute-force) மற்றும் 'நிறைவுவாதி' (perfectionist) என்று விவரிக்கப்பட்டது; முன்னதன் நோக்கம் எதிர்காலத்தில் கச்சா தரமான மொழிபெயர்ப்புகள் இருந்தால் பயனுள்ளதாக இருக்கும் ஒழுங்குமுறைகளின் உருவாக்கம்; பிந்தையதன் நோக்கம், மனிதர்களின் திருத்தம் சிறிதளவும் இல்லாத அல்லது தேவைப்படாத வெளியீட்டை உருவாக்கும் ஒழுங்குமுறைகளின் இறுதி வளர்ச்சி.

இந்த முதல் பத்தாண்டுகாலம் இயந்திர மொழிபெயர்ப்புக்கான மூன்று அடிப்படை அணுகுமுறைகளின் தொடக்கத்தைக் கண்டது (1980களின் பிற்பகுதியில் தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள் தோன்றும் வரை). முதலாவது 'நேரடி மொழிபெயர்ப்பு' மாதிரி ('direct translation model'); இதில் ஒரு மூல மொழியிலிருந்து (Source language (SL)) ஒரு குறிப்பிட்ட இலக்கு மொழிக்கு (target language (TL) குறைந்த அளவு பகுப்பாய்வு மற்றும் தொடரியல் மறுசீரமைப்புடன் மொழிபெயர்ப்பிற்காக நிரலாக்க விதிகள் உருவாக்கப்பட்டன. பல ஆராய்ச்சியாளர்கள் இருமொழி அகராதிகளை எளிதாக்கியும், அதாவது மூலமொழிச் சொற்களுக்கு ஒற்றை இலக்குமொழி நிகரன்களை வழங்குவதன் மூலம் ஒப்புருமொழிகள் (homonyms) மற்றும் பொருண்மைமயக்கச் சிக்கல்களைக் குறைக்க முயன்றனர்; இது பெரும்பாலான அர்த்தங்களை 'உள்ளடக்கும்'; எனவே சூழல்களின் (வழக்கமாக உடனடியாக அருகிலுள்ள சொற்கள்) பகுப்பாய்வைக் கோராது, மற்றும் முடிந்தவரை மூலமொழி மூலத்தின் சொல் வரிசையைப் பராமரிக்க அனுமதிக்கும்.

இரண்டாவது அணுகுமுறை அருவத்தன்மையான மொழி-நடுநிலை உருப்படுத்தங்களை (மூலமொழி மற்றும் இலக்குமொழி இரண்டிலிருந்தும் சுதந்திரமான குறியீடுகள் அல்லது சின்னங்கள்) அடிப்படையாகக் கொண்ட 'இடைமொழி' மாதிரி ('interlingua' model) ஆகும்; இதில் மொழிபெயர்ப்பு இரண்டு நிலைகளில் இருக்கும்: மூலமொழியிலிருந்து இடைமொழி மற்றும் இடைமொழியிலிருந்து இலக்குமொழி. மூன்றாவது அணுகுமுறை குறைந்த லட்சியம் கொண்டிருந்தது: 'இடமாற்ற அணுகுமுறை' ('transfer approach') என்பதில் இடமாற்ற நிலை வழியாக மூலப் பனுவல்களின் அருவத்தன்மையான (அதாவது பொருண்மை மயக்கம்

நீக்கப்பட்ட) உருப்படுத்தங்களில் இருந்தது நிகரான இலக்குமொழி உருப்படுத்தங்களுக்கு மாற்றம் நிகழும்; இந்த நேர்வில், மொழிபெயர்ப்பு மூன்று நிலைகள் உள்ளடக்கும்: பகுப்பாய்வு (analysis), இடமாற்றம் (transfer) மற்றும் உருவாக்கம் (generation) (அல்லது இணைப்பாக்கம் (synthesis)).

பெரும்பாலான சந்தர்ப்பங்களில், 'அனுபவவாதிகள்' (empiricists) பெரும்பாலும் அகராதி விதிகளைப் பெறுவதற்காக உண்மையான நூல்களின் புள்ளிவிவர பகுப்பாய்வுகளைப் பயன்படுத்தி 'நேரடி மொழிபெயர்ப்பு' (direct translation) அணுகுமுறையைப் பின்பற்றினர்; இது பெரும்பாலும் கோட்பாட்டு அடித்தளம் சிறிதளவு கொண்டோ அல்லது முற்றிலும் இல்லாமல் தற்காலிக இயல்புடையது, 'நிறைவு விரும்பிகள்' (perfectionists) வெளிப்படையாக கோட்பாட்டால் இயக்கப்பட்டவர்கள்; குறிப்பாகத் தொடரியல் பகுப்பாய்வு முறைகளுக்குக் கவனம் செலுத்தி அடிப்படை மொழியியல் ஆராய்ச்சியை மேற்கொண்டனர். சில குழுக்கள் இடைமொழி இலட்சியத்தைப் பின்தொடர்ந்தன; மற்றும் மனித சிந்தனை செயல்முறை குறித்த அடிப்படை ஆராய்ச்சி (பின்னர் செயற்கை நுண்ணறிவு என்று அல்லது புலனறிவு அறிவியல் அழைக்கப்படும்) மட்டுமே தானியங்கி மொழிபெயர்ப்பின் சிக்கல்களை தீர்க்கும் என்று நம்பப்பட்டது. மேலும் அவர்களில் நடைமுறைக்கேற்றவர்கள் தொடரியல் அடிப்படையிலான எளிமையான 'இடமாற்ற' மாதிரிகளில் (transfer models) கவனம் செலுத்தினர்; பொருண்மையியல் சிக்கல்களை சில பின்னர் நிலைக்கு விட்டுவிட்டனர்.

காலத்தின் எந்தவொரு மதிப்பீட்டும் கணினி வசதிகள் அடிக்கடி போதுமானதாக இல்லை என்பதை நினைவில் கொள்ள வேண்டும்; அடிப்படை வன்பொருளை மேம்படுத்துவதற்கு அதிக முயற்சி அர்ப்பணிக்கப்பட்டது (காகித நாடாக்கள், காந்த ஊடகங்கள், அணுகல் வேகம் போன்றவை) மற்றும் மொழி செயலாக்கத்திற்கு பொருத்தமான நிரலாக்கக் கருவிகளை உருவாக்குவது - குறிப்பாக, எம்ஐடி-இல் விக்டர் யங்வேயின் (Victor Yngve) கீழ் அணியால் COMIT உருவாக்கப்பட்டது. சில குழுக்கள் தவிர்க்க முடியாமல் கோட்பாட்டுப் பிரச்சினைகளில் கவனம் செலுத்த வேண்டிய கட்டாயம் ஏற்பட்டது, குறிப்பாக ஐரோப்பா மற்றும் சோவியத் யூனியனில். அரசியல் மற்றும் இராணுவ காரணங்களுக்காகக் கிட்டத்தட்ட அனைத்து யு.எஸ். ஆராய்ச்சிகளும் ரஷ்ய-ஆங்கில மொழிபெயர்ப்பிற்காக இருந்தது; மற்றும் பெரும்பாலான சோவியத் ஆராய்ச்சிகள் ஆங்கிலம்-ரஷ்ய மொழியில் கவனம் செலுத்தியன; இருப்பினும்

சோவியத் ஒன்றியத்தின் பன்மொழிக் கொள்கை அங்கு மற்ற இடங்களை விட மொழிகளின் பரந்த பரப்பில்லையில் ஆராய்ச்சிக்கு ஊக்கமளித்தது.

வாஷிங்டன் பல்கலைக்கழகத்தில் (Seattle/சியாட்டில்) எர்வின் ரீஃப்லரின் (Erwin Reifler) கீழ் மேற்கொள்ளப்பட்ட ஆராய்ச்சி அகராதி அடிப்படையிலான 'நேரடி' அணுகுமுறையின் (dictionary-based 'direct' approach) எடுத்துக்காட்டாக அமைந்தது; இது பெரிய இருமொழி அகராதிகளின் கட்டுமானத்தை உள்ளடக்கியது; இதில் அகராதிசார் தகவல்கள் சொல் நிகரன்களைத் தேர்ந்தெடுப்பதற்கு மட்டுமல்லாமல் தொடரியல் பகுப்பாய்வின் பயன்பாடு இல்லாமல் இலக்கணச் சிக்கல்களைத் தீர்க்கவும் பயன்படுத்தப்பட்டன. பதிவுகள் வெளியீட்டின் குறிப்பிட்ட இடம்சார் மறுநிரலாக்கதிற்கு (local reordering) விதிகளுடன் ஆங்கில மொழிபெயர்ப்புகளைக் கொடுத்தன. பிரமாண்டமான அகராதி, ரஷ்ய பல்பொருள் ஒருமொழிக்காக ஆங்கில 'உள்ளடக்குச் சொற்களை' விரிவாகப் பயன்படுத்துவது, தொடர்கள் மற்றும் எச்சத்தொடர்களை உட்படுத்துவது, மற்றும் சொற்றொகையைத் துணை மொழிகளாக வகைப்படுத்துவது இவற்றைச் செய்தது. ஜெர்மன் மற்றும் ஆங்கிலம் குறித்த ஆரம்ப வேலைகளுக்குப் பிறகு, இந்தக் குழு ரஷ்ய-ஆங்கிலத்தின் 'ஃபோட்டோஸ்கோபிக் ஸ்டோர்' (photoscopic store) என்ற ஒரு பெரிய நினைவக்க கருவியின் அஸ்திவாரங்களில் ஈடுபட்டது. 1958 முதல் ஐபிஎம் கார்ப்பரேஷனில் (யார்க்க்டவுன் ஹைட்ஸ், நியூயார்க் (Yorktown Heights, New York)) கில்பர்ட் கிங்கால் (Gilbert King) நடைமுறை வளர்ச்சி இயக்கப்பட்டது மற்றும் 1970களின் முற்பகுதி வரை அமெரிக்க விமானப்படைக்காக நிறுவப்பட்ட ஒழுங்குமுறை 'மொழிபெயர்ப்புகளை' தயாரித்தது. தரத்தில் வெளியீடு கச்சாவாகவும் சில நேரங்களில் அரிதாகவே புரியக்கூடியதாகவும் இருந்தது; ஆனால் ஒழுங்குமுறைக்கு அதிகப்படியான கோரிக்கை எதுவும் செய்யப்படவில்லை; இது அதன் எல்லாக் குறைபாடுடன் பயனர்களின் அடிப்படை தகவல் தேவைகளைத் திருப்தி செய்ய முடிந்தது.

அந்த நேரத்தில் பல ஆராய்ச்சியாளர்கள் மொழியியல் கோட்பாட்டில் அவநம்பிக்கை கொண்டனர் [ஜெல்லிங் ஹாரிஸின் (Zelling Harris) முறையான மொழியியல் (formal linguistics) மற்றும் நோம் சாம்ஸ்கி (Noam Chomsky) அரிதாகவே தொடங்கியிருந்தார்]; மேலும் மொழித் தரவுத்தொகுதிகளின் பகுப்பாய்வின் அடிப்படையில் ஆய்வு நெறிமுறைகளை உருவாக்க

விரும்பினார். எடுத்துக்காட்டாக, RAND கார்ப்பரேஷனின் ஆராய்ச்சியாளர்கள் இருமொழி பொருள்விளக்கச் சொற்கோவைகள் (bilingual glossaries) மற்றும் இலக்கண தகவல்களைப் பிரித்தெடுக்க ரஷ்ய இயற்பியல் நூல்களின் பெரிய தரவுத்தொகுதியின் புள்ளியியல்சார் பகுப்பாய்வுகளை மேற்கொண்டனர்; இதன் அடிப்படையில், ஒரு கணினி நிரல் ஒரு கரடுமுரட்டான மொழிபெயர்ப்பிற்காக எழுதப்பட்டது; இதன் விளைவு பின் திருத்தியமைப்பாளர்களால் ஆய்வு செய்யப்பட்டது; பொருள்விளக்கச் சொற்கோவைகள் (glossaries) மற்றும் விதிகள் திருத்தப்பட்டன; தரவுத்தொகுதி (corpus) மீண்டும் மொழிபெயர்க்கப்பட்டது; மொழிபெயர்ப்பு மற்றும் பிந்தைய எடிட்டிங் சுழற்சிகள் தொடர்ந்தது. பகுப்பாய்வின் முக்கிய நெறிமுறை ஆரம்பத்தில் புள்ளிவிவர விநியோகம் ஆகும்; RANDஇல் டேவிட் ஹேஸ் (David Hays) பின்னர் சார்பு இலக்கணத்தின் (dependency grammar) அடிப்படையில் முதல் தொடரியல் பாகுபடுத்தியை உருவாக்கினார்.

ஜார்ஜ்டவுன் பல்கலைக்கழகத்தில் (Georgetown University) லியோன் டோஸ்டெர்ட்டின் (Léon Dostert) கீழ் மேற்கொள்ளப்பட்ட ஆராய்ச்சி மிகவும் தேர்ந்தெடுக்கப்பட்ட அணுகுமுறையைக் (eclectic approach) கொண்டிருந்தது; பாரம்பரிய இலக்கணத் தகவல்கள் போதுமானதாக இல்லாதபோது மட்டுமே பனுவல்களின் அனுபவ பகுப்பாய்வுகளை (empirical analyses) மேற்கொள்வது. ஆரம்பத்தில் ஜார்ஜ்டவுனில் பல குழுக்கள் இருந்தன; இது பல ஆண்டுகளாக அமெரிக்காவில் மிகப்பெரிய குழுவாக இருந்தது. ஒரு குழு பால் கார்வினால் (Paul Garvin) வழி நடத்தப்பட்டது; பின்னர் பங்கர்-ராமோ கார்ப்பரேஷனில் (Bunker-Ramo Corporation) தனது சொந்தக் குழுவைக் கண்டுபிடிப்பதற்காக புறப்பட்டார்; அங்கு அவர் 'ஃபுல்க்ரம்' ('fulcrum') முறையை உருவாக்கினார்; அது அடிப்படையில் ஒரு சார்பு பாகுபடுத்தி ஆகும்; அரியட்னே லுக்ஜனோவ் தலைமையிலான மற்றொரு குழு குறியீடு பொருந்தும் முறையில் (code-matching method) பணியாற்றியது; மூன்றாவது ஒரு மனிதர் 'குழு' (ஆண்டனி பிரவுன் (Antony Brown)) ஒரு பிரெஞ்சு-ஆங்கில ஒழுங்குமுறைக்கு சுழற்சி முறையின் (cyclical method) தூய உதாரணத்துடன் பரிசோதனை செய்தார்; மற்றும் மைக்கேல் சரேக்னக்கின் (Michael Zarechnak) கீழ் நான்காவது குழு இறுதியில் பின்பற்றப்பட்ட முறையை உருவாக்கியது. இந்த ஜார்ஜ்டவுன் தானியங்கி மொழிபெயர்ப்பு (Georgetown Automatic Translation (GAT (கேட்)) ஒழுங்குமுறை

மூன்று நிலைப் பகுப்பாய்வுகளைக் கொண்டிருந்தது: உருபனியல் (மரபுமொழிகளை அடையாளம் காண்பது உட்பட), தொடரியல் (பெயர்ச்சொற்கள் மற்றும் பெயரடைகளின் உடன்பாடு, வினைச்சொற்களின் ஆளுமை, பெயரடைகளை மாற்றியமைத்தல் போன்றவை), மற்றும் தொடரியல் (எழுவாய்கள் மற்றும் பயனிலைகள், எச்சத்தொடர் உறவுகள் போன்றவை.) கேட் ஆரம்பத்தில் செர்னா (SERNA) ஒழுங்குமுறையில் செயல்படுத்தப்பட்டது; இது பெரும்பாலும் பீட்டர் டோமாவின் (Peter Toma) செயல்பாடாகும்; பின்னர் பிரவுன் (Brown) உருவாக்கிய நிரலாக்க முறை (programming method) மூலம் மேம்படுத்தப்பட்டது. இந்த வடிவத்தில் அது 1963ஆம் ஆண்டில் இஸ்ப்ராவில் (Ispa) (இத்தாலி (Italy) யூரடோம் (Euratom) ஆலும் 1964இல் அமெரிக்க அணுசக்தி ஆணையத்தாலும் (US Atomic Energey Commision) வெற்றிகரமாக நிறுவப்பட்டது; இரண்டும் 1970களின் பிற்பகுதி வரை வழக்கமான பயன்பாட்டில் தொடர்ந்தன.

ஹார்வர்ட் பல்கலைக்கழகத்தின் அந்தோனி ஓட்டிங்கர் (Anthony Oettinger) படிப்படியான அணுகுமுறையை (Gradualist approach) நம்பினார். 1954 முதல் 1960 வரை அவரது குழு ஒரு பெரிய ரஷ்ய-ஆங்கில அகராதியின் தொகுப்பில் கவனம் செலுத்தியது; இது மொழிபெயர்ப்பாளர்களுக்கு ஒரு கருவியாகவும் (இப்போது பொதுவான கணினி அடிப்படையிலான அகராதிக் கருவிகளின் முன்னோடி), விஷயத்தை நன்கு அறிந்த அறிவியலார்களுக்கு கச்சா சொல்லுக்குச்சொல் மொழிபெயர்ப்பை (crude word-for-word translation) உருவாக்கவும் மேலும் மேம்பட்டுத்தப்பட்ட பரிசோதனை செயல்பாடுகளுக்கு அடிப்படையாக இருக்கவும் தொகுக்கப்பட்டது. 1959ஆம் ஆண்டு முதல் ஆராய்ச்சி 'முன்கணிப்பு தொடரியல் பகுப்பாய்வி' ('predictive syntactic analyzer') உருவாக்கத்தில் கவனக்குவிப்பு செய்தது; முதலில் ஐடா ரோட்ஸின் (Ida Rhodes) கீழ் தேசிய தர நிர்ணய பணியகத்தில் (National Bureau of Standards) உருவாக்கப்பட்டது; இந்த ஒழுங்குமுறை இலக்கண வகைகளின் அனுமதிக்கப்பட்ட வரிசைமுறைகளை (பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைகள் போன்றவை)அடையாளம் காண்பதற்கும் பின்வரும் வகைபாடுகளின் நிகழ்தகவுசார் கணிப்புக்கும் உதவும். இருப்பினும், பெரும்பாலும் திருப்தியற்றதாக இருந்தன; இதன் முதன்மைக் காரணம் 'மிகவும் சாத்தியமான' கணிப்பின் ஒவ்வொரு கட்டத்திலும் செயல்படுத்தப்பட்ட தேர்வு முடிவுகள் ஆகும். (ஆயினும்கூட, மேம்படுத்தப்பட்ட பதிப்பான பன்முக-வழி முன்கணிப்பு பகுப்பாய்வி

(Mutiple-path Predictive Analyzer), பின்னர் வில்லியம் உட்ஸின் (William Woods') பழக்கமான பெரிதாக்கப்பட மாற்ற வலைப்பின்னல் பாகுபடுத்திக்கு (Augmented Transition Network parser) வழிவகுத்தது.)

எம்ஐடியில் (MIT) ஆராய்ச்சி 1953முதல் 1965ஆம் ஆண்டு அதன் இறுதி வரை விக்டர் யங்வே-ஆல் (Victor Yngve) இயக்கப்பட்டது. இதில் தொடரியல் மையத்தில் வைக்கப்பட்டது: ஒரு மூலமொழி இலக்கணமானது உள்ளீட்டு வாக்கியங்களை தொடரமைப்பு உருப்படுத்தங்களாக (phrase structure representations) பகுப்பாய்வு செய்தது; 'கட்டமைப்பு மாற்ற வழக்கம்' ('structure transfer routine') அவற்றை நிகரான இலக்குமொழி தொடரமைப்புகளாக மாற்றியது; இலக்குமொழி இலக்கண விதிகள் வெளியீட்டு உரையை உருவாக்கியது. ஆனால் இறுதியில், ஒரு "சொற்பொருண்மை தடை" ("semantic barrier") எட்டப்பட்டிருப்பதையும் மேலும் முன்னேற்றம் மிகவும் கடினமாக இருக்கும் என்பதையும் யங்வே உணர்ந்தார். ஒரு குறுகிய காலத்திற்குக் குழுவுக்கு சாம்ஸ்கியின் தொடர்பு இருந்தபோதிலும் மாற்றிலக்கணம் (transformational grammar) செல்வாக்கைக் கொண்டிருக்கவில்லை; அச்சமயத்தில் எந்த இயந்திரமொழிபெயப்பு ஆராய்ச்சியிலும் சாம்ஸ்கிசார் அணுகுமுறைகளின் சான்றுகள் உண்மையில் கிட்டத்தட்ட இல்லை.

1958ஆம் ஆண்டில் வின்ஃபிரைட் லெஹ்மன் (Winfried Lehmann), டெக்சாஸ் பல்கலைக்கழகத்தில் (University of Texas) நிறுவிய மொழியியல் ஆராய்ச்சி மையம் (Linguistic Research Centre (LRC)), அடிப்படை தொடரியல் ஆராய்ச்சியிலும் கவனம் செலுத்தியது. அடிப்படையில் 'தொடரியல் மாற்றம்' (syntactic transfer) அணுகுமுறையில் இரு திசை மொழிபெயர்ப்பை (bi-directional translation) அடைய, தலைமாற்ற இலக்கணங்களை (reversible grammars) உருவாக்க முயற்சிகள் மேற்கொள்ளப்பட்டன; இது பிற்காலத்தில் மெட்டல் (METAL) ஒழுங்குமுறையின் வளர்ச்சிக்கு அதற்கான அடித்தளங்களை அமைத்தது.

பெர்க்லியில் (Berkeley) உள்ள கலிபோர்னியா பல்கலைக்கழகத்தில் (University of California) சிட்னி லாம்பின் (Sydney Lamb) வழிகாட்டுதலின் கீழ் ஆய்வுத் திட்டம் அதிகபட்ச திறமையான அகராதி நடைமுறைகளை உருவாக்குவதன் முக்கியத்துவத்தையும் இயந்திர மொழிபெயர்ப்புக்குப் பொருத்தமான மொழியியல் கோட்பாட்டையும் வலியுறுத்தியது. இது கணினிகளின் கட்டமைப்புக்கு இணையாக உள்ளன வலைப்பின்னகள், கணுக்கள் மற்றும்

உறவுகள் கொண்ட லாம்பின் அடுக்கமைவு இலக்கணமாகும் (Stratificational Grammar). மொழிபெயர்ப்பு, தொடர்ச்சியான அடுக்குகள் (series of strata) (எழுத்துசார், உருபனியல்சார், சொல்சார், பொருண்மைசார்) வழியாக குறியத்திறவு (encoding) மற்றும் குறியனாக்கச் (coding) செயல்முறைகளின் தொடர்ச்சிகளாகக் கருதப்பட்டது.

இடைமொழி அணுகுமுறையை (interlingua approach) ஏற்கும் அமெரிக்க குழுக்கள் எதுவும் இல்லை; வீவரின் (Weaver's) முந்தைய வாதத்தை மீறி அமெரிக்க திட்டங்கள் குறைவான ஊக அணுகுமுறைகளை (speculative approaches) ஏற்றுக்கொள்ள முனைந்தன. இடைமொழி அணுகுமுறைகள் (Intelinguas) வேறு திட்டங்களின் கவனக்குவிப்பாக இருந்தது. கேம்பிரிட்ஜ் மொழி ஆராய்ச்சி பிரிவில் (Cambridge Languge Research Unit), மார்கரெட் மாஸ்டர்மேன் (Margaret Masterman) மற்றும் அவரது சகாக்களும் இரண்டு அடிப்படை ஆராய்ச்சி வழிகளைப் பின்பற்றினர்: கச்சா 'பிட்ஜின்' (அடிப்படையில் வார்த்தைக்கு வார்த்தை) மொழிபெயர்ப்புகளை உருவாக்கும் இடைமொழி மூலமுன்மாதிரியின் உருவாக்கம்; மற்றும் முதன்மையாக ஒரு சொற்களஞ்சியத்தின் (thesaurus) வளமான பெருண்மையியல் வலைப்பின்னல்கள் (semantic networks) மூலம் (அடிப்படையாக அணிக்கோவையைப் (lattice) பயன்படுத்துகிற கோட்பாடு) இயந்திர மொழிபெயர்ப்பு வெளியீட்டை மேம்படுத்துவதற்கும் சீராக்குவற்கும் கருவிகளின் உருவாக்கம். மிலனில் (Milan), சில்வியோ செக்காடோ (Silvio Ceccato) 'புலனறிவு' செயல்முறைகள் (cognitive processes) அடிப்படையில் ஒரு இடைமொழியின் உருவாக்கத்தில் கவனம் செலுத்தினார்; குறிப்பாக சொற்களின் கருத்துருசார் பகுப்பாய்வு (conceptual analysis) (இனங்கள், பேரினம், செயல்பாட்டு வகை, இயற்பியல் பண்புகள், போன்றன) மற்றும் பனுவல்களில் உள்ள பிற சொற்களுடன் அவற்றின் சாத்தியமான தொடர்புகள்; இதன் முன்னோடி பிற்காலத்தின் 'நரம்பியல் வலைப்பின்னகள்' ('neural networks') ஆகும்.

சோவியத் யூனியனில் ஆராய்ச்சி அமெரிக்காவைப் போலவே தீவிரமானது மற்றும் இதேபோன்ற அனுபவ மற்றும் கோட்பாட்டு அணுகுமுறைகள் கலவையைக் காட்டியது. பிரிசிஸன் மெக்கானிக்ஸ் நிறுவனத்தில் (Institute of Precision Mechanics) டி.ஒய். பனோவ் (D.Y. Panov) என்பவரின் கீழ் ஆராய்ச்சி ஆங்கிலம்-ரஷ்ய மொழிபெயர்ப்பில் ஜார்ஜ்டவுனில் இருந்ததைப் போன்ற வழிகளில் இருந்தது; ஆனால் முதன்மையாக போதுமான அளவு கணினி

வசதிகள் இல்லாததால் நடைமுறை வெற்றி குறைவாக இருந்தது. அதிக அளவிலான அடிப்படை ஆராய்ச்சி ஸ்டெக்லோவ் கணித நிறுவனத்தில் (Steklov Mathematical Institute) அலெக்ஸேஜ் ஏ. லஜபுனோவ் (Aleksej A. Ljapunov), ஓல்கா எஸ். குலகினா (Olga S. Kulagina) மற்றும் இகோர் ஏ. மெல்'யுக் (Igor A. Mel'čuk (மொழியியல் நிறுவனத்தின்); பிந்தையவர் ஒரு இடைமொழிய அணுகுமுறையில் பணியாற்றினார்; இது அவரது 'பொருள்-உரை' மாதிரிக்குக் ('meaning-text' model) கொண்டு சென்றது. இது அடுக்கமைவுசார் சார்பு அணுகுமுறையை (stratificational dependency approach) (ஆறு அடுக்கு: ஒலியியல்சார், ஒலியனியல்சார், உருபனியல்சார், புற அமைப்புத் தொடரியல்சார், அக அமைப்புத் தொடரியல்சார், பொருண்மையியல்சார்) இடைமொழிய அணுகுமுறையின் அகராதிசார் அம்சங்களின் வலுவான முக்கியத்துவத்துடன் இணைத்தது. அடுக்கு உறவுகளை (paradigmatic relations) உள்ளடக்கிய ஆழமான தொடரியல்சார் அடுக்கில் (deep syntactic stratum) ஐம்பது உலகளாவிய 'சொல்சார் செயல்பாடுகள்' (lexical functions) அடையாளம் காணப்பட்டன (எ.கா. ஒருபொருள்பன்மொழிகள் (synonyms), எதிர்மொழிகள் (antonyms), வினைச்சொற்கள் மற்றும் அவற்றின் தொடர்புடைய செயலிசார் பெயர்ச்சொற்கள் (agentive nouns) போன்றன) மற்றும் பலவகையான தொடரியல் உறவுகள் [எ.கா. கொடுக்கப்பட்ட பெயர்ச்சொற்களுடன் தொடர்புடைய தொடக்கம்சார் வினைச்சொற்கள் (inceptive verbs), conference: open, war: break out); மரபுத்தொடர்சார் இயக்குவினைகள் (idiomatic causatives), compile: dictionary, lay: foundation போன்றன].

இடைமொழிசார் ஆய்வுகள் சோவியத் ஒன்றியத்தின் பன்மொழி தேவைகளுடன் உடன்பட்டிருக்கின்றது; பல மையங்களில் மேற்கொள்ளப்பட்டது (ஆர்க்கைம்பால்ட் மற்றும் லியோன் (Archaimbault and Léon) 1997). முதன்மையான ஒன்று லெனின்கிராட் மாநில பல்கலைக்கழகத்தில் (Leningrad State University) இருந்தது; அங்கு நிகோலாஜ் ஆண்ட்ரீவின் (Nikolaj Andreev) கீழ் ஒரு குழு ஒரு இடைமொழியை அருவத்தன்மையான இடைப்பட்ட உருப்படுத்தமாகக் கருதவில்லை; ஆனால் முழு செயற்கை மொழியாக அதன் சொந்த உருபனியல் மற்றும் தொடரியல் உடனும் புள்ளியியல் அடிப்படையில் பல்மொழிகளுக்குப் பொதுவான பண்புக்கூறுகளை மட்டுமே கொண்டிருந்தது.

1960களின் நடுப்பகுதியில் இயந்திரமொழிபெயர்ப்பு ஆராய்ச்சி குழுக்கள் பெரும்பாலான ஐரோப்பிய நாடுகளை உள்ளடக்கி (ஹங்கேரி, செக்கோஸ்லோவாக்கியா, பல்கேரியா, பெல்ஜியம், ஜெர்மனி, பிரான்ஸ் (Hungary, Czechoslovakia, Bulgaria, Belgium, Germany, France), முதலியன), சீனா, மெக்சிகோ மற்றும் ஜப்பான் (China, Mexico, and Japan) பல நாடுகளில் நிறுவப்பட்டன. இவற்றில் பல குறுகிய காலமே இருந்தன; ஒரு விதிவிலக்கு 1960இல் கிரெனோபில் பல்கலைக்கழகத்தில் தொடங்கப்பட்ட திட்டம் ஆகும் (பார்க்க கீழே பிரிவு 2.5.).

இந்த காலகட்டம் முழுவதும், இயந்திரமொழிபெயர்ப்பு பற்றிய ஆராய்ச்சி அமைப்பு மொழியியல் (structural linguistics) மற்றும் முறையான மொழியியல் (formal linguistics) (குறிப்பாக சோவியத் ஒன்றியத்தில்), குறியீட்டியல், தருக்கப் பொருண்மையியல் (logical semantics), கணித மொழியியல் (mathematical linguistics), அளவு மொழியியல் (quantitative linguistics) மற்றும் இப்போது கணினி மொழியியல் (computational linguistics) மற்றும் மொழி பொறியியல் (language engineering) (1960களின் முற்பகுதியில் இருந்து ஏற்கனவே பயன்பாட்டில் உள்ள சொற்கள்) என அழைக்கப்படும் கிட்டத்தட்ட அனைத்தும் மிகவும் சமகாலச் செயல்பாடுகளுக்கு ஒரு 'குடை' ஆனது. ஆரம்பத்தில், தன்னாள்வியல் (சைபர்நெடிக்ஸ் (cybernetics)) மற்றும் தகவல் கோட்பாட்டுடன் (information theory) நெருங்கிய உறவுகளும் இருந்தன. பொதுவாக, ஆரம்ப காலம் முழுவதும், இயந்திர மொழிபெயர்ப்பு (கோட்பாட்டு மற்றும் நடைமுறை இரண்டும்) தொடர்பான பணிகள் சம்பந்தப்பட்ட பல துறைகளில் 'அறிவுசார்' பணிகளுக்குக் கணினிகளைப் பயன்படுத்துவதன் மூலம் பரவலான பொருத்தமாக காணப்பட்டன. இது குறித்த ஆராய்ச்சிக்கு குறிப்பாக உண்மை இயந்திர மொழிபெயர்பின் 'இடைமொழி' அம்சங்கள், ஆவண மீட்டெடுப்பு அமைப்புகளில் பயன்படுத்தப்படுவதற்கு 'தகவல் மொழிகளின்' உருவாக்கத்திற்கு முக்கியத்துவம் வாய்ந்ததாகக் கருதப்படுகின்றது.

2.4. அல்பாக்அறிக்கை மற்றும் அதன் விளைவுகள்

ALPAC அறிக்கை மற்றும் அதன் விளைவுகள் (The ALPAC report and its consequences) என்ற தலைப்பில் ஹட்சின்ஸ் (Hutchins 2015) கூறியுள்ள செய்திகள் இத்தலைப்பில் தொகுத்துத் தரப்பட்டுள்ளன. 1950களில் நம்பிக்கை அதிகமாக இருந்தது; கணினி மற்றும் முறையான மொழியியலில் முன்னேற்றங்கள், குறிப்பாகத் தொடரியல் பகுதியில் தரத்தில் சிறந்த

மேம்பாடுகளை உறுதிப்படுத்துவதாகத் தோன்றியது. உடனடி முன்னேற்றங்கள் மற்றும் சில ஆண்டுகளில் முழு தானியங்கி ஒழுங்குமுறைகள் செயல்படும் கணிப்புகள் பல இருந்தன. இருப்பினும், மொழியியல் சிக்கல்களின் கலவைத்தன்மை மேலும் மேலும் தெளிவாகத் தெரிந்ததால் ஏமாற்றம் அதிகரித்தது, வெளிப்படையாக ஈடுசெய்ய முடியாத 'பொருண்மையியல்சார் தடையை' ('semantic barrier') ஆராய்ச்சி அடைந்துவிட்டதாக பலர் ஒப்புக்கொண்டனர். ஒரு செல்வாக்கான கணக்கெடுப்பில் பார்-ஹில்லெல் (Bar-Hillel 1960) மனித மொழிபெயர்ப்பாளர்களிடமிருந்து முடிவுகளை பிரித்தறிய முடியாத வகையில் முழுமையான தானியங்கி உயர் தர மொழிபெயர்ப்பு (fully automatic high quality translation (FAHQT)) அமைப்புகளை உருவாக்குதல் இயந்திரமொழிபெயர்ப்பின் ஆராய்ச்சியின் குறிக்கோளாக இருக்க வேண்டும் என்ற நிலவும் கருத்தை விமர்சித்தார். மொழியியல் அறிவு மற்றும் கணினி அமைப்புகள் தற்போதைய நிலையைப் பொறுத்தவரை இது வெறுமனே நம்பத்தகாதது அல்ல, ஆனால் கொள்கையளவில் சாத்தியமற்றது என்று அவர் வாதிட்டார். அவர் pen என்ற வார்த்தையுடன் தனது வாதத்தை நிரூபித்தார். இது குறைந்தது இரண்டு அர்த்தங்களைக் கொண்டிருக்கலாம் (விலங்குகள் அல்லது குழந்தைகளுக்கான ஒரு கொள்கலன், மற்றும் ஒரு எழுதும் கருவி). The box was in the pen என்ற வாக்கியத்தில் முதல் பொருள் மட்டுமே நம்பத்தகுந்ததாக நமக்குத் தெரியும்; இரண்டாவது பொருள் பேனாக்கள் மற்றும் பெட்டிகளின் சாதாரண அளவுகள் பற்றிய நமது அறிவால் விலக்கப்படுகிறது. பார்-ஹில்லெல் இத்தகைய சிக்கலான எடுத்துக்காட்டுகள் பொதுவானவை என்றும் எந்தவொரு கணினி நிரலும் பரந்த கலைக்களஞ்சிய சேகரிப்பின் உதவியைநாடாமல் இதுபோன்ற 'உண்மையான உலக' அறிவைப் புரிந்து கையாள இயலாது என்றும் வாதிட்டார். அவரது வாதம் அந்நேரத்தில் அதிக மதிப்பைக் கொண்டிருந்தது, இருப்பினும் செயற்கை நுண்ணறிவின் (artificial intelligence) வளர்ச்சிகள் (மற்றும் இயந்திரமொழிபெயர்ப்பில் அறிவு அடிப்படையிலான ஒழுங்குமுறைகள், பிரிவு 2.7 கீழே) அவரது அவநம்பிக்கை முற்றிலும் நியாயம் இல்லை என்பதை நிரூபித்துள்ளது.

சில காலமாக இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சி தீவிரமாக தொடர்ந்தது; உண்மையில் பல புதிய குழுக்கள் குறிப்பாக யுனைடெட் ஸ்டேட்ஸ் மற்றும் ஐரோப்பாவிற்கு வெளியே அமைக்கப்பட்டன; ஆராய்ச்சி அப்போது முக்கியமாக தொடரியல் பகுப்பாய்விலும்

பொருண்மையியலின் தொடக்க ஆய்விலும் கவனம் செலுத்தியது. அதே நேரத்தில், முதல் வேலைசெய்யும் ஒழுங்குமுறைகள் (ஐபிஎம் மற்றும் ஜார்ஜ்டவுன்-இல் இருந்து) நிறுவப்பட்டன; பல ஆராய்ச்சியாளர்களின் பார்வையில் அவை முன்கூட்டியே (முதிராத நிலையில்) நிறுப்பட்டதாகும்; மற்றும் உடனடி முடிவுகளை விரும்பும் பயனர்களால் மோசமான-தரமான மொழிபெயர்ப்புகள் கிடைப்பது பாராட்டப்பட்டது; அவர்களுக்கு மனிதத் தரமான (human-quality) பதிப்புகள் தேவையில்லை.

ஆயினும்கூட, நல்ல தரமான இயந்திர மொழிபெயர்பின் உடனடி வாய்ப்பு குறைந்து கொண்டிருந்தது; 1964இல் அமெரிக்காவில் இயந்திர மொழிபெயர்பின் அரசாங்க ஆதரவாளர்கள் (முக்கியமாக இராணுவ மற்றும் புலனாய்வு அமைப்புகள்) நிலைமையை ஆராய தானியங்கி மொழி செயலாக்க ஆலோசனைக் குழுவை (Automatic Language Processing Advisory Committee (ALPAC) அமைக்க தேசிய அறிவியல் நிறுவனம் (National Science Foundation) கேட்டனர். அதன் புகழ்பெற்ற 1966 அறிக்கையில், இயந்திர மொழிபெயர்ப்பு மெதுவான, குறைவான துல்லியமான மற்றும் மனித மொழிபெயர்ப்பை விட இரண்டு மடங்கு விலை உயர்ந்தது என்றும், "உடனடி அல்லது கணிக்கக்கூடியது எதுவுமில்லை பயனுள்ள இயந்திர மொழிபெயர்ப்பின் வாய்ப்பு இல்லை" (ALPAC 1966) என்றும் முடிவு கூறப்பட்டது. இது இயந்திர மொழிபெயர்ப்பு ஆய்வில் மேலும் முதலீடு செய்யத் தேவையில்லை என்று கண்டது; அதற்கு பதிலாக மொழிபெயர்ப்பாளர்களுக்கான தானியங்கி அகராதிகள் (automatic dictionaries) போன்ற இயந்திர துணைக்கருவிகளை (machine aids) உருவாக்கவும் கணினி மொழியியலில் அடிப்படை ஆராய்ச்சியின் தொடர்ச்சியான ஆதரவையும் பரிந்துரைத்தது. முரண்பாடாக, ALPAC இயந்திர மொழிபெயர்ப்பை நிராகரித்தது; ஏனெனில் அதற்குப் பிந்தைய திருத்தியமைப்பு (post-editing) தேவைப்பட்டது (மனித மொழிபெயர்ப்புகளும் கூட வெளியீட்டிற்கு முன்னர் மாற்றியமைக்கப்பட்டது) மற்றும் உயர்தர மொழிபெயர்ப்புகள் தேவை என்று அது கருதியது. ஆனால் நிதி ஆதரவு அமைப்புகள் முதன்மையாக தகவல் சேகரிப்பு மற்றும் பகுப்பாய்வில் ஆர்வமாக இருந்தன; அவற்றிற்கு குறைந்த தரம் ஏற்றுக்கொள்ளத்தக்கதாக இருந்தது. அந்த நேரத்தில் ஒரு பக்கச்சார்பான மற்றும் குறுகிய பார்வை கொண்டதாக பரவலாக கண்டனம் செய்யப்பட்டாலும், ALPACஇன் செல்வாக்கு ஆழமானது; இது அமெரிக்காவில் இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சிக்கு ஒரு பதின்ம ஆண்டுகளுக்கு மேல் மெய்நிகர் முடிவைக் கொடுத்தது; மறைமுகமாக வேறு இடங்களிலும் இயந்திர மொழிபெயர்பு ஆராய்ச்சியை முடிவுக்குக் கொண்டுவந்தது. சோவியத்தில் நிதியளிக்கும் அமைப்புகள் மிகவும் ஏழ்மையான கணினி வசதிகளுடன் வெற்றிக்கான வாய்ப்புகள் இன்னும் சிறியவை என்று வாதாடியது. மேலும்

ALPAC அறிக்கை, கணினிகள் மற்றும் இயற்கை மொழியின் ஆய்வில் இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியின் முன்னணி பகுதி ஆகும் என்ற முந்தைய கருத்தை முடிவுக்குக் கொண்டுவந்தது. கணினி மொழியியல் ஒரு சுதந்திர துறையாக மாறியது.

2.5.மூன்றாவது காலகட்டம் (1967-1976)

ஹட்சின்ஸ் (Hutchins 2015) இக்காலகட்டத்தை "அமைதியான பதின்ம ஆண்டு' (The quiet decade) என்று குறிப்பிடுகின்றார். ALPAC அறிக்கைக்குப் பின்னரும் ஆராய்ச்சி முழுமையாக நிறுத்தப்படவில்லை. அமெரிக்காவில் கூட ஒரு சில குழுக்கள் டெக்சாஸ் பல்கலைக்கழகத்திலும், வெய்ன் மாநில பல்கலைக்கழகத்திலும் (Wayne State University) மேலும் ஆண்டுகள் தொடர்ந்தன. ஆனால் திசையில் மாற்றம் ஏற்பட்டது. ALPAC க்கு முந்தைய காலத்தின் (1956-1966) "முதல் தலைமுறை" ஆராய்ச்சி முக்கியமாக 'நேரடி மொழிபெயர்ப்பு' அணுகுமுறைகளால் ('direct translation' approaches) ஆதிக்கம் செலுத்தப்பட்டது; ALPACக்கு பிந்தைய (post-ALPAC) "இரண்டாம் தலைமுறை" ("second generation") இடைமொழிசார் (interlingua) மற்றும் இடமாற்றம்சார் (transfer) அடிப்படையிலான 'மறைமுக' ('indirect') மாதிரிகளால் ஆதிக்கம் செலுத்தப்பட வேண்டும்.

யுனைடெட் ஸ்டேட்ஸில் முக்கிய செயல்பாடு ரஷ்ய மொழியின் அறிவியல் மற்றும் தொழில்நுட்ப ஆணவங்களின் ஆங்கில மொழிபெயர்ப்புகளில் கவனம் செலுத்தியது. கனடா மற்றும் ஐரோப்பாவில் தேவைகள் முற்றிலும் வேறுபட்டவை. கனடியன் அரசாங்கத்தின் இரு கலாச்சாரக் கொள்கை மொழிபெயர்ப்புத் தொழிலின் திறனைத் தாண்டி ஆங்கிலம்-பிரஞ்சுக்கான (மற்றும் குறைந்த அளவிற்கு பிரெஞ்சு-ஆங்கிலம்) மொழிபெயர்ப்பு கோரிக்கையை உருவாக்கியது. மொழிபெயர்ப்பின் சிக்கல்கள் ஐரோப்பிய சமூகத்திற்குள் விஞ்ஞான, தொழில்நுட்ப, நிர்வாக மற்றும் சட்ட ஆவணங்கள் எல்லா சமூக மொழிகளிலிருந்தும் மற்றும் எல்லா சமூக மொழிகளுக்கும் மொழிபெயர்ப்புகளுக்கான வளர்ந்து வரும் கோரிக்கைகளுடன் சமமான கடுமையில் இருந்தன. யுனைடெட் ஸ்டேட்டில் இயந்திர மொழிபெயர்ப்பு பல ஆண்டுகளாகப் புத்துயிர் பெறாமல் இருக்கும்போது கனடாவிலும் ஐரோப்பாவிலும் (பின்னர்

ஜப்பானிலும் பிற இடங்களிலும்) இயந்திர மொழிபெயர்பின் தேவை அங்கீகரிக்கப்படுவது நிறுத்தப்படவில்லை; வளர்ச்சி தொடர்ந்தது.

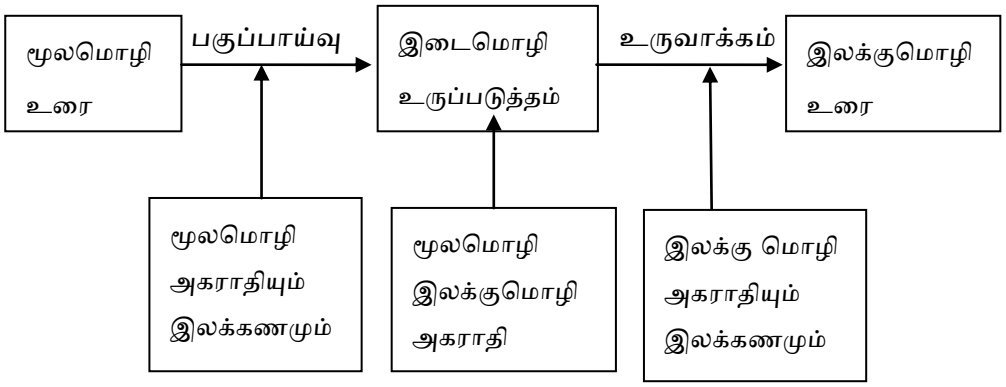
மாண்ட்ரீலில் (Montreal) ஆங்கிலம்-பிரஞ்சு மொழிபெயர்ப்பிற்கான ஒரு தொடரியல் இடமாற்ற ஒழுங்குமுறை (syntactic transfer system) குறித்து 1970இல் ஆராய்ச்சி தொடங்கியது. TAUM திட்டம் (Traduction Automatique de l'Université de Montréal) இரண்டு முக்கிய சாதனைகளைக் கொண்டிருந்தது: முதலாவதாக, மொழியியல் கோர்வைளையும் கிளையமைப்புகளையும் கையாளுவதற்கான க்யூ-ஒழுங்குமுறை வடிவவாதம் (Q-system formalism) (பின்னர் புரோலாக் நிரலாக்க மொழி என உருவாக்கப்பட்டது); இரண்டாவதாக, வானிலை முன்னறிவிப்புகளை மொழிபெயர்ப்பதற்கான மெட்டோ ஒழுங்குமுறை (Météo system). குறிப்பாக வானிலை அறிக்கைகளின் கட்டுப்படுத்தப்பட்ட சொற்றொகை மற்றும் வரையறுக்கப்பட்ட தொடரியல் ஆகியவற்றிற்காக வடிவமைக்கப்பட்ட மெட்டியோ 1976 முதல் வெற்றிகரமாக இயங்கியது (1984 முதல் புதிய பதிப்பில்). TAUM குழு இந்த வெற்றியை விமான கையேடுகளின் மற்றொரு துணை மொழியுடன் மீண்டும் செய்ய முயற்சித்தது; ஆனால் கலவைத்தன்மையான பெயர்ச்சொல் கூட்டுகள் மற்றும் தொடர்களின் சிக்கல்களை சமாளிக்க தவறிவிட்டது; திட்டம் 1981இல் முடிவடைந்தது.

பதின்ம ஆண்டின் முக்கிய புதுமையான சோதனைகள் இன்றியமையாததாக இடைமொழி அணுகுமுறைகளில் கவனம் செலுத்தியது. 1960க்கும் 1971க்கும் இடையில் கிரெனோபிள் பல்கலைக்கழகத்தில் (Grenoble University) பெர்னார்ட் வாகோயிஸால் (Bernard Vauquois) நிறுவப்பட்ட குழு ரஷ்ய கணிதம் மற்றும் இயற்பியல் பனுவல்களைப் பிரெஞ்சு மொழியில் மொழிபெயர்க்கும் ஒரு ஒழுங்குமுறை உருவாக்கப்பட்டது. அதன் 'மைய மொழி' (pivot language) (ரஷ்யாவில் குலஜினா மற்றும் மெல்'யுக் (Kulagina and Mel'čuk) ஆராய்ச்சி மூலம் ஓரளவிற்கு செல்வாக்கு செலுத்தியது) தொடரியல் உறவுகளின் தர்க்கரீதியான பண்புகளை மட்டுமே உருப்படுத்தம் செய்தது; இது ஒரு தூய்மையான இடைமொழி அல்ல, ஏனெனில் இது சொற்களுக்கு ஒன்றோடொன்று உருப்படுத்தங்களை வழங்கவில்லை; இவை இருமொழி இடமாற்ற இயந்திரநுட்பத்தால் (bilingual transfer mechanism) மொழிபெயர்க்கப்பட்டன. பகுப்பாய்வு மற்றும் உருவாக்கம் மூன்று நிலைகள் சம்பந்தப்பட்டது: தொடர் அமைப்பு (சூழல்

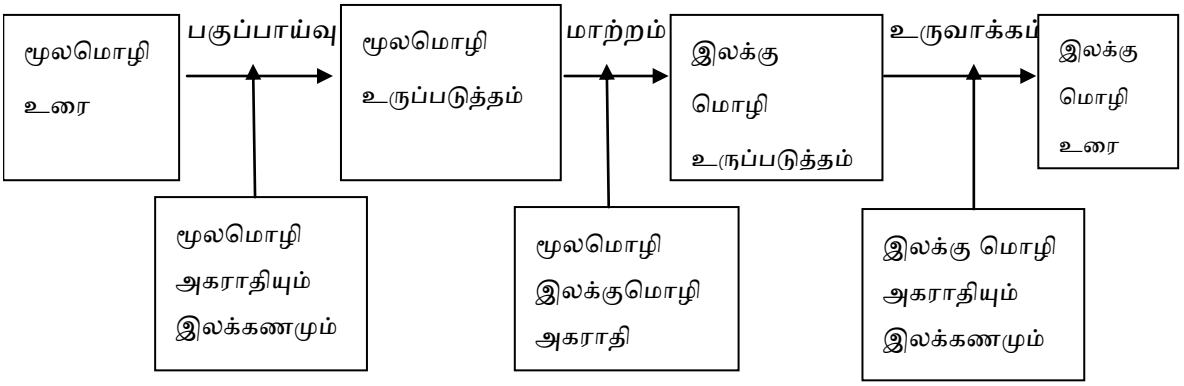
கட்டுப்படில்லாத) (phrase-structure (context-free)) உருப்படுத்தம், சார்பு அமைப்பு (dependency structure) மற்றும் பனிலை மற்றும் பங்கேற்பாளகள் (predicates and arguments) அடிப்படையில் 'மைய மொழி' உருப்படுத்தம். டெக்சாஸ் பல்கலைக்கழகத்தில் 1970களில் ஜெர்மன் மற்றும் ஆங்கிலத்திற்கான மெட்டல் அமைப்பில் இதேபோன்ற மாதிரி ஏற்றுக்கொள்ளப்பட்டது: வாக்கியங்கள் 'இயல்பான வடிவங்கள்' ('normal forms') ஆகப் பகுப்பாய்வு செய்யப்பட்டன; அதாவது இடைமொழி சொல்சார் கூறுகள் இல்லாத பொருண்மையியல்சார் கருத்துரை சார்பு கட்டமைப்புகள் (semantic propositional dependency structures). அதே நேரத்தில் சோவியத் யூனியனில் மெல்'யுக் இயந்திர மொழிபெயர்ப்பு பயன்பாட்டிற்கு வேண்டி ஒரு 'பொருள்-உரை' மாதிரி ('meaning-text' model) குறித்த தனது ஆராய்ச்சியைத் தொடர்ந்தார் (மேலே காண்க).

இருப்பினும், 1970களின் நடுப்பகுதியில் இடைமொழிசார் அணுகுமுறையின் எதிர்காலம் சந்தேகத்திற்குரியதாகத் தோன்றியது. அடையாளம் காணப்பட்ட முக்கிய சிக்கல்கள் கிரெனோபிள் மற்றும் டெக்சாஸ் குழுக்களால் (Grenoble and Texas groups) பகுப்பாய்வு நிலைகளின் கடினத்தன்மைக்கு காரணமாக இருந்தன (எந்த ஒரு கட்டத்திலும் தோல்வி என்பது எந்தவொரு வெளியீட்டையும் உற்பத்தி செய்யத் தவறியது): பாகுபடுத்திகளின் திறமையின்மை (வடிகட்டப்பட வேண்டிய பல பகுதி பகுப்பாய்வுகள்), குறிப்பாக மூல மொழியின் உள்ளீட்டின் புற வடிவங்கள் பற்றிய தகவல் இழப்பு, இலக்குமொழி வடிவங்களையும் ஏற்றுக்கொள்ளக்கூடிய இலக்குமொழி வாக்கிய அமைப்புகளின் கட்டுமானத்தையும் தேர்ந்தெடுப்பதற்கு வழிகாட்டப் பயன்படுத்தப்பட்டிருக்கலாம். இதன் விளைவாக, குறைந்த லட்சிய 'இடமாற்ற' அணுகுமுறை சிறந்த வாய்ப்புகளை வழங்கியது என அந்த நேரத்தில் பலருக்குத் தோன்றியது.

பின்வரும் படம் இடைமொழியைப் பயன்படுத்து ஒரு எடுத்துக்காட்டான ஒழுங்குமுறையை விளக்கும்.



பின்வரும் படம் பரிமாற்றம்/இடமாற்றைத்தை உள்ளடக்கிய இடைமொழிவாயிலான இயந்திரமொழிபெயர்ப்பின் மூன்று மட்ட நிலையை உருப்படுத்தம் செய்யும்.



2.6. நான்காவது காலகட்டம் (1976-1989)

இக்காலகட்டம் ஹட்சின்ஸால் (Hutchins 2015) "செயல்பாட்டு மற்றும் வணிக ஒழுங்குமுறைகள்" **(Operational and commercial systems)** என்று குறிப்பிடப்பட்டுள்ளது. ALPACக்குப் பிந்தைய பதின்ம ஆண்டில் அதிகமான ஒழுங்குமுறைகள் செயல்பாட்டு பயன்பாட்டிற்கு வந்து பொதுமக்களை கவனம் ஈர்த்தன. ஜார்ஜ்டவுன் ஒழுங்குமுறைகள்

1960களின் நடுப்பகுதியில் இருந்து இயங்கி வந்தன. அத்துடன் மெட்டியோ, மற்றும் பிற இரண்டு துணை மொழி ஒழுங்குமுறைகள் (sublanguage systems) தோன்றின: 1970ஆம் ஆண்டில் இன்ஸ்டிட்யூட் டெக்ஸ்டைல் டி பிரான்ஸ் (Institut Textile de France) ஒரு பன்மொழி ஒழுங்குமுறையான டைடஸ்-ஐ (TITUS) அறிமுகப்படுத்தியது; மற்றும் 1972 இல் சீன ஹாங்காங் பல்கலைக்கழகம் (Chinese University of Hong Kong (CULT)) குறிப்பாக சீன மொழியிலிருந்து கணித நூல்களை ஆங்கிலத்தில் மொழிபெயர்க்க வடிவமைக்கப்பட்டது.

இருப்பினும், மிகவும் முக்கியமானது, முதல் சிஸ்ட்ரான் நிறுவல்கள் (Systran installations). பீட்டர் டோமா (Peter Toma) உருவாக்கிய அதன் பழமையான பதிப்பு, யுஎஸ்ஏஎஃப் வெளிநாட்டு தொழில்நுட்ப பிரிவில் (USAF Foreign Technology Division) (Dayton டேட்டன், Ohio/ஓஹியோ) 1970இல் நிறுவப்பட்ட ரஷ்ய-ஆங்கில ஒழுங்குமுறை ஆகும். ஐரோப்பிய சமூகங்களின் ஆணையம் (Commission of the European Communities) ஒரு ஆங்கில-பிரஞ்சு பதிப்பை 1976இல் வாங்கியது மற்றும் ஐரோப்பிய சமூகங்களின் (European Communities) (இப்போது ஐரோப்பிய ஒன்றியம் (European Communities)) பிற மொழிகளின் மொழிபெயர்ப்பிற்கான ஒழுங்குமுறைகள் அதைத் தொடர்ந்தது. பல ஆண்டுகளாக அதிகரித்த கூறுநிலைமை (modularity) மற்றும் வேறுபட்ட புதிய மொழி இணைகளை உருவாக்கும்போது செலவுக் குறைப்புகளை அனுமதிக்கிற வேறுபட்ட பதிப்புகளின் பகுப்பாய்வு மற்றும் கூட்டிணைப்பாகத்தின் அதிகப் பொருந்தக்கூடிய தன்மை (compatibility) இவற்றால் அசல் ('நேரடி மொழிபெயர்ப்பு') வடிவமைப்பு பெரிதும் மாற்றப்பட்டுள்ளது. சிஸ்ட்ரான் ஏராளமான அரசுகளுக்கிடையிலான நிறுவனங்களில் (intergovernmental institutions) நிறுவப்பட்டது; எ.கா. நேட்டோ மற்றும் சர்வதேச அணுசக்தி ஆணையம் (NATO and the International Atomic Energy Authority), மற்றும் பல பெரிய நிறுவனங்களில், எ.கா. ஜெனரல் மோட்டார்ஸ் (General Motors), டோர்னியர் (Dornier) மற்றும் ஏரோஸ்பேட்டியேல் (Aérospatiale). ஜெராக்ஸ் கார்ப்பரேஷனில் (Xerox Corporation) குறிப்பாக அதன் பயன்பாடு குறிப்பிடத்தக்கதாக இருந்தது: ஆங்கிலத்திலிருந்து பிரஞ்சு, ஜெர்மன், இத்தாலியன், ஸ்பானிஷ், போர்த்துகீசியம் மற்றும் ஸ்காண்டிநேவிய மொழிகளுக்கு (French, German, Italian, Spanish, Portuguese, and Scandinavian languages) மொழிபெயர்ப்பதற்கு தொழில்நுட்ப கையேடுகளின் சொற்றொகையையும் கட்டமைப்புகளையும் கட்டுப்படுத்துவதால் பிந்தைய திருத்த அமைப்பு மூலம் கிட்டத்தட்ட அகற்றப்பட்டது.

1980களின் முற்பகுதியிலிருந்து சமீபம் வரை சிஸ்ட்ரானின் (Systran) முக்கிய போட்டியாளர் லோகோ கார்ப்பரேஷனலிருந்து (Logos Corporation) வந்த ஒழுங்குமுறை,ஆகும்; இது ஆரம்பத்தில் பெர்னார்ட் ஈ. ஸ்காட் (Bernard E. Scott) என்பவரால் விமான கையேடுகளை

மொழிபெயர்ப்பதற்கான ஆங்கில-வியட்நாமிய ஒழுங்குமுறையாக (English-Vietnamese system) 1970களில் உருவாக்கப்பட்டது. இத்திட்டத்தில் பெறப்பட்ட அனுபவம் 1982இல் சந்தையில் தோன்றிய ஜெர்மன்-ஆங்கில ஒழுங்குமுறையின் (German-English system) வளர்ச்சிக்கு பயன்படுத்தப்பட்டது; 1980களில் பிற மொழி இணைகள் உருவாக்கப்பட்டன.

1980களின் இறுதியில் வணிகரீதியான மெட்டல் ஜெர்மன்-ஆங்கில (METAL German-English system) முறை தோன்றியது; இது டெக்சாஸ் பல்கலைக்கழக ஆராய்ச்சியில் இருந்து உருவானது. 1970களின் நடுப்பகுதியில் அதன் இடைமொழிசார் சோதனைகளுக்குப் பிறகு இக்குழு அடிப்படையில் இடமாற்ற அணுகுமுறையை பின்பற்றியது; இந்த ஆராய்சிக்கு 1978 முதல் முனிச்சில் (Munich) (ஜெர்மனி) சீமென்ஸ் நிறுவனம் (Siemens company) நிதியளித்தது. பிற மொழி இணைகள் பின்னர் டச்சு, பிரஞ்சு மற்றும் ஸ்பானிஷ் மற்றும் ஆங்கிலம் (Dutch, French and Spanish) மற்றும் ஜெர்மன் (English and German).

சிஸ்ட்ரான் (Systran), லோகோஸ் (Logos) மற்றும் மெட்டல் (METAL) போன்ற ஒழுங்குமுறைகள் கொள்கை அடிப்படையில் பொதுவான பயன்பாட்டிற்காக வடிவமைக்கப்பட்டன; இருப்பினும் நடைமுறையில் அவற்றின் அகராதிகள் குறிப்பிட்ட பொருள் களங்களுக்கு மாற்றியமைக்கப்பட்டுள்ளன. ஒரு குறிப்பிட்ட சூழலுக்காக வடிவமைக்கப்பட்ட சிறப்பு-தேவை ஒழுங்குமுறைகள் 1970கள் மற்றும் 1980களில் உருவாக்கப்பட்டன. வாஷிங்டனில் உள்ள பான் அமெரிக்கன் ஹெல்த் ஆர்கனைசேஷன் (Pan American Health Organization in Washington) இரண்டு மெயின்பிரேம் (mainframe) ஒழுங்குமுறைகளை உருவாக்கியது; ஒன்று ஸ்பானிஷ் மொழியிலிருந்து ஆங்கிலம் (SPANAM) மற்றும் மற்றொன்று ஆங்கிலத்திலிருந்து ஸ்பானிஷ் மொழி (ENGSPAN); இரண்டுமே அடிப்படையில் முரியல் வாஸ்கான்செலோஸ் (Muriel Vasconcellos) மற்றும் மார்ஜோரி லியோன் (Marjorie León) என்ற இரண்டு ஆராய்ச்சியாளர்களால் உருவக்கப்பட்டன. பெரிய குறிப்பிட்ட நோக்கத்திற்கான ஒழுங்குமுறைகள் (tailor-made systems) 1980களின் முற்பகுதியில் இருந்து ஸ்மார்ட் கார்ப்பரேஷனின் (Smart Corporation) (நியூயார்க்) சிறப்பாகும். வாடிக்கையாளர்கள் சிட்டிகார்ப், ஃபோர்டு (Citicorp, Ford) மற்றும் எல்லாவற்றிலும் மிகப்பெரிய கனேடிய வேலைவாய்ப்பு மற்றும் குடிவரவுத் துறை (Canadian Department of Employment and Immigration) இவற்றை உட்படுத்தும். ஸ்மார்ட் ஒழுங்குமுறைகளின் (Smart systems) (ஜெராக்ஸைப் (Xerox) போல) முதன்மையான

அம்சம் வெளியீட்டின் திருத்தத் தேவையைக் குறைக்கும் படி உள்ளீட்டு (ஆங்கிலம்) சொற்றொகை மற்றும் தொடரியலின் கண்டிப்பான கட்டுப்பாடு ஆகும்.

1980களில், மிகப் பெரிய வணிக செயல்பாடு ஜப்பானில் இருந்தது, அங்கு பெரும்பாலான கணினி நிறுவனங்கள் (புஜித்சூ, ஹிட்டாச்சி, என்.இ.சி, ஷார்ப், தோஷிபா (Fujitsu, Hitachi, NEC, Sharp, Toshiba)) கணினி உதவி மொழிபெயர்ப்பிற்கான மென்பொருளை முக்கியமாக ஜப்பானிய-ஆங்கிலம் மற்றும் ஆங்கிலம்-ஜப்பானிய சந்தைகளுக்கு (Japanese-English and English-Japanese markets) உருவாக்கியன; இருப்பினும் அவை கொரிய, சீன (Korean, Chinese) மற்றும் பிற மொழிகளுக்கு மொழிபெயர்ப்பு தேவைகளைப் புறக்கணிக்கவில்லை. இந்த ஒழுங்குமுறைகளில் பல மட்டுப்படுத்தப்பட்ட உருபனியல் மற்றும் தொடரியல் தகவல்களுக்குப் பகுப்பாய்வு செய்கிற மற்றும் சொற்பொருண்மை மயக்கங்களை நீக்கச் சிறிய அளவில் முற்றிலும் முயற்சி செய்யாத தாழ் நிலை நேரடி அல்லது இடமாற்ற ஒழுங்குமுறைகள் ஆகும். பெரும்பாலும் குறிப்பிட்ட பாடத் துறைகளுக்கு (கணினி அறிவியல் மற்றும் தகவல் தொழில்நுட்பம் என்பன பிரபலமான விருப்பத்தேர்வுகள்), அவை தயாரிப்பு (முன் திருத்தம் (pre-editing)) மற்றும் திருத்தம் (பின் திருத்தம் (post-editing) நிலைகள் இரண்டிலும் கணிசமான மனித உதவியை நம்பியிருந்தன.

சில ஜப்பானிய அமைப்புகள் மைக்ரோ கம்ப்யூட்டர்களுக்காக (microcomputers) வடிவமைக்கப்பட்டன. ஆனால் அவைகள் இந்த சந்தையில் முதன்மையானது இல்லை. ஆரம்பகாலத்தியன முறையே 1981இல் அமெரிக்க வீட்னர் (American Weidner) மற்றும் 1983இல் ALPS ஒழுங்குமுறைகள். ALPS ஒழுங்குமுறை மூன்று நிலை உதவிகளை வழங்கியது: பன்மொழி சொல் செயலாக்கம் (word-processing), தானியங்கி அகராதி (automatic dictionary) மற்றும் கலைச் சொல் ஆலோசனை (terminology consultation) மற்றும் ஊடாடும் மொழிபெயர்ப்பு (interactive translation). பிந்தைய நேர்வில் இயந்திரமொழிபெயர்ப்பு மூலம் உருவாக்கப்பட்ட செப்பமில்லா மொழிபெயர்ப்பு வரைவுகள் (MT-produced rough drafts) மீது மொழிபெயர்ப்பாளர்கள் பணியாற்றலாம். இந்த ஒழுங்குமுறை 'மொழிபெயர்ப்பு நினைவகம்' ('translation memory') என்பதன் தொடக்க வடிவத்தைக் கொண்டிருந்தது (கீழே பிரிவு 2.8 ஐப் பார்க்கவும்.) இருப்பினும், ALPS தயாரிப்புகள் லாபகரமானவை அல்ல; 1980 களின் நடுப்பகுதியில் இருந்து இந்நிறுவனம் மொழிபெயர்ப்பாளர்களுக்குக் கணினி பயன்பாட்டுக் கருவிகள் விற்பதை விட மொழிபெயர்ப்பு

சேவையை வழங்குவதில் திசை திருப்பப்பட்டது. வீட்னர் ஒழுங்குமுறைகள் ஏராளமான மொழி இணைகளுக்கு தொகுப்புகளை வழங்கின; அதன் ஜப்பானிய-ஆங்கில ஒழுங்குமுறைகள் குறிப்பாக பிரபலமானவை. 1980களின் பிற்பகுதியில் வீட்னர் (Weidner) பிராவிஸால் (Bravice) கையகப்படுத்தப்பட்டது; ஆனால் விரைவில் நிறுவனம் முடிவுக்கு வந்தது. இருப்பினும், இந்த நேரத்தில் தனிப்பட்ட கணினிகளுக்கான பிற ஒழுங்குமுறைகள் சந்தையில் வந்துவிட்டன (மொழியியல் தயாரிப்புகளிலிருந்து (Linguistic Products) பிசி-மொழிபெயர்ப்பாளர் (PC-Translator), குளோபலிங்கிலிருந்து (Globalink) ஜி.டி.எஸ் (GTS) மற்றும் மைக்ரோடாக்கிலிருந்து (MicroTac) மொழி உதவித் தொடர் (Language Assistant series).

2.7. ஐந்தாவது கட்டம் (1976 முதல் 1989 வரை)

ஹட்சின்ஸ் (Hutchins 2015) இதை "ஆராய்ச்சியின் மறுமலர்ச்சி" (The revival of research) என்று குறிப்பிடுகின்றார். 1970களின் பிற்பகுதியிலும் 1980களின் முற்பகுதியிலும் மொழிபெயர்ப்பு ஆராய்ச்சியின் மறுமலர்ச்சி மூன்று கட்ட இடமாற்ற அடிப்படையிலான அணுகுமுறையை (transfer-based approach) முக்கியமாக தொடரியல் சார்ந்த அணுகுமுறையை ஏறக்குறைய உலகளாவிய ரீதியில் ஏற்றுக்கொள்வதன் மூலம் பண்பாக்கம் செயப்பட்டது மற்றும் அன்றைய காலத்திய மொழியிய கோட்பாடுகளால் பாதிக்கப்பட்டுள்ள சொல்சார் மற்றும் இலக்கண விதிகளை முறைப்படுத்துவதால் நிறுவப்பட்டது.

கிரெனோபிள் குழு (கேதா (GETA (Groupe d'Etudes pour la Traduction Automatique)) அதன் இடைமொழிய ஒழுங்குமுறையின் ஏமாற்றத்திற்குப் பிறகு, அதன் செல்வாக்குமிக்க அரியேன் ஒழுங்குமுறையின் (Ariane system) உருவாக்கதைத் தொடங்கியது. "இரண்டாம் தலைமுறை" ("second generation") மொழியியல் அடிப்படையிலான இடமாற்ற ஒழுங்குமுறைகளின் (linguistics-based transfer systems) முன்னுதாரணம் எனக் கருதப்பட்ட அரியேன் 1980களில் உலகம் முழுவதும் திட்டங்களைப் பாதித்தது. குறிப்பாக அதன் நெகிழ்வுத்தன்மை (flexibility) மற்றும் கூறுநிலைமை (modularity, கிளை அமைப்பு உருப்படுத்தங்களைக் கையாளும் அதற்கான வழிமுறைகள் மற்றும் நிலையான மற்றும் மாறும் இலக்கணங்களின் கருத்து ஆகியவை குறிப்பிடத்தக்கவை. வெவ்வேறு நிலைகளும் உடுப்படுத்தத்தின் வகைகளும் (சார்பு, தொடரமைப்பு, தருக்கம் (dependency, phrase structure, logical)) ஒற்றை வகை புலக்குறிப்பு செய்யப்பட்ட கிளை அமைப்புகளால் (labelled tree structures) இணைக்கப்படலாம்; மற்றும் இதனால் பன்னிலை இடமாற்ற உருப்படுத்தங்களில் (multilevel

transfer representations) கணிசமான நெகிழ்வுத்தன்மையை வழங்குகிறது. எனினும், பல சோதனை இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளைப் போலவே அரியேன் ஒரு செயல்பாட்டு ஒழுங்குமுறையாக (operational system) மாறவில்லை (ஒரு பிரெஞ்சு தேசிய இயந்திர மொழிபெயர்ப்பு திட்டத்தில் ஈடுபாடு இருந்தபோதிலும்); மற்றும் 1980களின் பிற்பகுதியில் இந்த ஒழுங்குமுறை குறித்த செயல்பாடான ஆராய்ச்சி நிறுத்தப்பட்டது.

கெத்தா அரியனே (GETA-Ariane) வடிவமைப்பை ஒத்ததான மு (Mu) ஒழுங்குமுறை மாகோடோ நாகோவின் (Makoto Nagao) கீழ் கியோட்டோ பல்கலைக்கழகத்தில் (University of Kyoto) உருவாக்கப்பட்டது. மு-இன் முக்கிய அம்சங்கள் வேற்றுமை இலக்கண பகுப்பாய்வு மற்றும் சார்பு கிளையமைப்பு (dependency tree) உருப்படுத்தங்களின் பயன்பாடு மற்றும் இலக்கணம் எழுதுவதற்கு (GRADE) நிரலாக்க சூழலின் உருவாக்கம். கியோட்டோ ஆராய்ச்சி பல ஜப்பானிய இயந்திர மொழிபெயர்ர்பு ஆராய்ச்சி திட்டங்களிலும் பல ஜப்பானிய வணிக ஒழுங்குமுறைகளிலும் (Japanese commercial systems) பெரும் தாக்கத்தை ஏற்படுத்தியது. 1986 முதல் ஆராய்ச்சி முன்மாதிரி ஜப்பானிய அறிவியல் மற்றும் தொழில்நுட்பத்திற்கான தகவல் மையத்தால் (Japanese Information Centre for Science and Technology) சுருக்கங்களை மொழிபெயர்க்கப் பயன்படுத்த ஒரு செயல்பாட்டு ஒழுங்குமுறையாக மாற்றப்பட்டது.

சர்ப்ரூக்கனில் (Saarbrücken) (ஜெர்மனி) சோதனை ஆராய்ச்சி 1967இல் தொடங்கியது; 1970களின் நடுப்பகுதியில் இருந்து நுணுக்கமுறைகளின் பன்முகத்தன்மையைக் காட்டுகிற ஒரு பன்மொழி பரிமாற்ற ஒழுங்குமுறை SUSY (Saarbrücker Übersetzungssystem) உருவாக்கப்பட்டது: தொடரமைப்பு விதிகள் (phrase structure rules), மாற்றம்சார் விதிகள் (transformational rules), வேற்றுமை இலக்கணம் (case grammar) மற்றும் இணைதிறன் சட்டங்கள் (valency frames), சார்பு இலக்கணம் (dependency grammar), and புள்ளியியல்சார் தரவுகளின் (statistical data) பயன்பாடு போன்றவை. இதன் முக்கிய கவனம் ரஷ்ய மொழி மற்றும் ஜெர்மன் மொழி போன்ற திரிபுறும் மொழிகளின் (inflected languages) ஆழமான ஆய்வாகும்; ஆனால் ஆங்கிலம் மற்றும் பிரஞ்சு உள்ளிட்ட பிற மொழிகளும் ஆராயப்பட்டன. இந்தக் குழு ஜப்பானிய அறிவியல் கட்டுரைகளின் தலைப்புகளை ஜெர்மன் மொழியில் மொழிபெயர்க்கும் பொருட்டு புஜித்சு அட்லாஸ் ஒழுங்குமுறைலிருந்து (Fujitsu ATLAS system) வெளியீட்டை மாற்றுவதற்காக ஒரு உருவாக்கியை (SEMSYN) உருவாக்கியது.

1980களில் நன்கு அறியப்பட்ட திட்டங்களில் ஒன்று ஐரோப்பிய சமூகங்களின் யூரோட்ரா திட்டம் (Eurotra project) ஆகும். அதன் நோக்கம் அனைத்து சமூக மொழிகளுக்குள்ளும் மொழிபெயர்ப்பிற்கான மேம்பட்ட பன்மொழிய பரிமாற்ற முறையை உருவாக்குவதாகும்; இது சமூகங்களின் சிஸ்ட்ரான் ஒழுங்குமுறையின் 'நேரடி மொழிபெயர்ப்பு' அணுகுமுறை இயல்பாகவே மட்டுப்படுத்தப்பட்டிருந்தது என்ற அனுமானத்தின் அடிப்படையில் ஆகும். கெத்தா அரியனே (GETA-Ariane) மற்றும் ஸூஸி (SUSY) போன்று இதன் வடிவமைப்பு பன்னிலை இடைமுகங்களில் அதிக அளபிலான அருவத்தன்மையாக்கத்தில் சொல்சார் (lexical), தர்க்க-தொடரியல் (logico-syntactic) மற்றும் பொருண்மையியல் தகவல்களை (semantic information) இணைத்துள்ளது. கூடுதல் மொழியியல் அறிவுத் தளங்களின் அல்லது அனுமான இயந்திர நுட்பத்தின் பயன்பாடு எதுவும் செய்யப்படவில்லை; மற்றும் மொழிபெயர்ப்பு செயல்முறைகளின் போது மனித உதவி அல்லது தலையீட்டிற்கான வசதிகள் எதுவும் இணைக்கப்படவில்லை. சம்பந்தப்பட்டவர்களால் உடனடியாக ஒப்புக்கொள்ளப்பட்ட ஒரு பெரிய குறைபாடு, கோட்பாட்டு அடிப்படையிலும் நடைமுறை அடிப்படையிலும் அகராதியின் சிக்கல்களைச் சமாளிக்கத் தவறியது, இந்தத் திட்டம் சமூகம் முழுவதிலும் பல பல்கலைக்கழக ஆராய்ச்சி குழுக்களை உள்ளடக்கியது; ஆனால் 1980களின் முடிவில் எந்தவொரு செயல்பாட்டு முறையும் எதிர்பார்க்கப்படவில்லை மற்றும் திட்டம் முடிந்தது, இருப்பினும் கணினி மொழியியலில் குறுக்கு தேசிய ஆராய்ச்சியைத் (cross-national research) தூண்டும் அதன் இரண்டாம் நோக்கத்தை அடைந்தது.

1980களின் பிற்பகுதியில் செயற்கை நுண்ணறிவு மற்றும் புலறிவு மொழியியலில் சமகால ஆராய்ச்சி மூலம் ஓரளவு ஊக்கப்படுத்தப்பட்ட இடைமொழிய ஒழுங்குமுறைகளில் ஆர்வத்தின் பொதுவான மறுமலர்ச்சி இருந்தது. விநியோகிக்கப்பட்ட மொழி மொழிபெயர்ப்பு (Distributed Language Translation (DST) ஒழுங்குமுறை உட்ரெக்டில் (Urecht) (The Netherlands (நெதர்லாந்து)) உள்ள பிஎஸ்ஓ (BSO) மென்பொருள் நிறுவனத்தில் டூன் விட்காமின் (Toon Witkam) வழிகாட்டுதலின் கீழ் கணினி வலிப்பின்னகளின் (computer networks) மேல் செயல்படும் ஒரு பன்மொழிய ஊடாடும் ஒழுங்குமுறையாக (multilingual interactive system) கருதப்பட்டது; இதில் ஒவ்வொரு முனையமும் ஒரு மொழியில் இருந்து மட்டுமே மொழிபெயர்க்கும் இயந்திரமாக

இருக்க வேண்டும். உரைகள் எஸ்பரேன்டோவின் (Esperanto) மாற்றியமைக்கப்பட்ட வடிவமான இடைநிலை மொழியில் (Intermiday) உள்ள முனையங்களுக்கு (terminal) இடையில் அனுப்பப்பட வேண்டும். பகுப்பாய்வு முதன்மையாக உருபனியல் மற்றும் தொடரியல் பண்புக்கூறுகளுடன் (சார்பு இலக்கணத்தில் முறைபடுத்தப்பட்டது) கட்டுப்படுத்தப்பட்டது; பொருண்மையியல் செயலாக்கம் இல்லை; பொருண்மை மயக்கநீக்கம் மைய இடைமொழியக்கூறில் (interlingua component) நடந்தது. ஆய்வுத்திட்டம் பெரிய சொல்சார் தரவுத்தளங்களின் கட்டுமானத்தில் ஒரு குறிப்பிடத்தக்க முயற்சியை மேற்கொண்டது; மற்றும் அதன் இறுதி ஆண்டுகளில் (மனித) மொழிபெயர்க்கப்பட்ட பனுவல்களின் (Sadler 1989) தரவுத்தொகுதியில் இருந்து ஒரு இருமொழி அறிவு வங்கியை (Bilingual Knowledge Bank) உருவாக்க முன்மொழியப்பட்டது; இந்த விஷயத்தில் பிற்கால எடுத்துக்காட்டு அடிப்படையிலான ஒழுங்குமுறைகளை (example-based systems) எதிர்பார்க்கிறது (பார்க்க பிரிவு 2.9 கீழே).

மற்றொரு விஷயத்தில் புதுமையான நெதர்லாந்தில் (Netherlands) இரண்டாவது இடைமொழியத் திட்டம், பிலிப்ஸில் (Philips) (ஐன்ட்ஹோவன் (Eindhoven)) ஜான் லாண்ட்ஸ்பெர்கன் (Jan Landsbergen) இயக்கிய ரொசெட்டா (Rosetta) திட்டம் ஆகும். மாண்டேக் இலக்கணத்தின் (Montague grammar) இடைமொழி உருப்படுத்தங்களில் பயன்பாட்டை ஆராய்வதே இதன் நோக்கம்; கூட்டமைவு கொள்கையைப் (principle of compositionality) பின்பற்றி பொருண்மையியல்சார் உருப்படுத்தங்கள் வெளிப்பாடுகளின் தொடரியல் அமைப்பில் இருந்து ஆக்கப்பட்டன; ஒவ்வொரு தொடரியல்சார் ஆக்கக் கிளையமைப்பிற்கும் (syntactic derivation tree) அதனுடன் தொடர்புடைய பொருண்மையியல் ஆக்கக் கிளையமைப்பு (semantic derivation tree) இருக்க வேண்டும்; மேலும் இந்த பொருண்மையியல்சார் ஆக்கக் கிளை அமைப்புகள் இடைமொழி உருப்படுத்தங்கள் ஆகும். இரண்டாவது முக்கியமான அம்சம், இலக்கணங்களின் தலைகீழ்மாற்றம் (reversibility) ஆராய்வது; அதாவது ஒரு மொழியின் தொடரியல் மற்றும் பொருண்மையியல் பகுப்பாய்விற்காக ஒரு திசையில் செயல்படும் இலக்கண விதிகள் மற்றும் மாற்றங்களின் தொகுப்பு; மறு திசையில் அந்த மொழியின் சரியான வாக்கியங்களின் உருவாக்கம் (உற்பத்தி). தலைகீழ்மாற்றம் பல அடுத்தடுத்த இயந்திர மொழிபெயர்ப்பு திட்டங்களின் அம்சமாக மாறியது.

1980களின் பிற்பகுதியில் ஜப்பான் அதன் இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியில் கணிசமான அதிகரிப்பைக் கண்டது. பெரும்பாலான கணினி நிறுவனங்கள் [புஜித்சூ (Fujitsu), தோஷிபா (Thoshiba), ஹிட்டாச்சி (Hitachi), போன்றவை] பெரிய தொகைகளை முதலீடு செய்யத்

தொடங்கின; இதை அரசாங்கமும் தொழில்துறையும் வரவிருக்கும் தகவல் சமூகத்தின் "ஐந்தாவது தலைமுறைக்கு" ("fifth generation") அடிப்படையாகக் கருதியது. கியோட்டோ பல்கலைக்கழகத்தில் (Kyoto University) மு திட்டத்தால் (Mu project) ஆரம்பத்தில் பெரிதும் பாதிக்கப்பட்டுள்ள ஆராய்ச்சி, பலவிதமான அணுகுமுறைகளைக் காட்டியது. பரிமாற்ற ஒழுங்குமுறைகள் ஆதிக்கம் செலுத்தியபோது, இடைமொழிய ஒழுங்குமுறைகளும் இருந்தன எ.கா. நெக்-இல் (NEC) உள்ள பிவட் (PIVOT) ஒழுங்குமுறை மற்றும் 1980களின் நடுப்பகுதியில் (மற்றும் இன்றுவரை தொடர்கிறது) சீனா (China), இந்தோனேசியா (Indonesia), மலேசியா (Malaysia) மற்றும் தாய்லாந்து (Thailand) இவற்றின் பங்கேற்புடனும் மற்றும் முக்கிய ஜப்பானிய ஆராய்ச்சி நிறுவனங்களின் ஈடுபாடுடனும் தொடங்கப்பட்ட ஜப்பானிய நிதியுதவி செய்த பன்மொழி பன்னாட்டு திட்டம் (multilingual multinational project).

1980களில் பல ஆராய்ச்சி திட்டங்கள் வட அமெரிக்காவிற்கு, மேற்கு ஐரோப்பா, மற்றும் ஜப்பான் இவற்றிற்கு வெளியே அமைக்கப்பட்டன - கொரியாவில் (சில நேரங்களில் ஜப்பானிய மற்றும் அமெரிக்க குழுக்களுடன் கூட்டு திட்டங்களில்), தைவானில் (எ.கா. ஆர்க்ட்ரான் ஒழுங்குமுறை (ArchTran system)), சீனாவின் பிரதான நிலப்பரப்பில் பல நிறுவனங்களில், மற்றும் தென்கிழக்கில்ஆசியா, குறிப்பாக மலேசியாவில். சோவியத் யூனியனில் செயல்பாட்டில் அதிகரிப்பு இருந்தது. 1976 முதல் பெரும்பாலான ஆராய்ச்சிகள் மாஸ்கோவில் உள்ள அனைத்து யூனியன் மொழிபெயர்ப்பு மையத்திலும் குவிக்கப்பட்டன. ஆங்கிலம்-ரஷ்யன் (ஆம்பர் (AMPAR)) மற்றும் ஜெர்மன்-ரஷ்ய மொழிபெயர்ப்பு (நெர்பா (NERPA)) ஒழுங்குமுறைகள் நேரடி அணுகுமுறை அடிப்படையில் உருவாக்கப்பட்டன; ஆனால் மெல்'யுக்கின் (Mel'čuk's) 'அர்த்தம் உரை' (meaning-text) மாதிரி அடிப்படையில் யூரிஜ் அப்ரெஸ்ஜானின் (Yurij Apers'jan) வழிகாட்டுதலின் கீழ் செயல்பாடு இருந்து - 1977இல் சோவியத் யூனியனை விட்டு வெளியேற மெல்யூக் கடமைப்பட்டிருந்தார். இது மேம்பட்ட பரிமாற்ற ஒழுங்குமுறைகள் (advanced transfer systems) ஃப்ராப் (FRAP) (பிரெஞ்சு-ரஷ்ய மொழிக்கு), மற்றும் எடாப் (ETAP) (ஆங்கிலம்-ரஷ்ய மொழிக்கு) வழிவகுத்தது. இருப்பினும் இந்த குழு தவிர சோவியத் ஒன்றியத்தின் பெரும்பாலான நடவடிக்கைகள் ஒப்பீட்டளவில் குறைந்த-நிலை செயல்பாட்டு ஒழுங்குமுறைகளின் பெரும்பாலும் புள்ளியியல்சார் பகுப்பாய்வுகளின் பயன்பாட்டை உள்ளடக்கிய உற்பத்தியில் கவனம் செலுத்தியது - அங்கு லெனின்கிராட் மாநில பல்கலைக்கழகத்தில் (Leningrad State University) ரைமண்ட் பியோட்ரோவ்ஸ்கியின் (Raimund Piotrowski) கீழ் 'பேச்சின் புள்ளிவிவரக் குழுவின் ('Speech Statistics' group) செல்வாக்கு குழு ரஷ்யாவில் பல பிற்கால வணிக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் உருவாக்கத்திற்குக் குறிப்பாகக் குறிப்பிடத்தக்கதாக உள்ளது.

1980களில் பல ஆராய்ச்சியாளர்கள் இயந்திர மொழிபெயர்ப்பு தரத்தை மேம்படுத்துவதற்கான வழிமுறைகள் செயற்கை நுண்ணறிவு (Artificial Intelligence (AI)) சூழலில் இயற்கையான மொழி ஆய்வு ஆராய்ச்சியிலிருந்து வரும் என்று நம்பினர். இயந்திர மொழிபெயர்ப்பில் செயற்கை நுண்ணறிவு முறைகளின் ஆய்வுகள் 1970களின் நடுப்பகுதியில் யோரிக் வில்க்ஸின் (Yorick Wilks) 'விருப்பத்தேர்வு பொருண்மையியல்' (preference semantics) 'மற்றும் 'பொருண்மையியல் வார்ப்புருக்கள்' (semantic templates) [அதாவது இருப்பவைகள் (entities), கருத்துருக்கள் (concepts), எழுவாய்-வினை (subject-verb), வினை-நேரடிச்செயப்படுபொருள் (verb-direct object) போன்ற குறிப்பிட்ட கட்டமைப்பு உறவுகளில் உள்ள செயல்கள் (activities) என்பனவற்றின் மிகவும் பொதுவான அல்லது மிகவும் விருப்பமான சேர்ந்துவருகைகளை (collocations) அடையாளம் காண்பதற்கான வழிகள்] தொடர்பான பணிகளுடன் தொடங்கியது. யேல் பல்கலைக்கழகத்தில் (Yale University) ரோஜர் ஷாங்க் (Roger Schank) ஆராய்ச்சியிலிருந்து, குறிப்பாக பனுவல் 'புரிதலுக்கு' ('understanding') நிபுணர் ஒழுங்குமுறைகள் (expert systems) மற்றும் அறிவு-அடிப்படையிலான அணுகுமுறைகளின் (knowledge-based approaches) உருவாக்கத்திலிருந்து மேலும் உத்வேகம் வந்தது.

பல திட்டங்கள் அறிவு சார்ந்த அணுகுமுறைகளைப் பயன்படுத்தின - ஜப்பானில் சில (எ.கா. NTTஇல் LUTE திட்டம், மற்றும் ஜப்பானிய பன்மொழி திட்டத்திற்கான ETL ஆராய்ச்சி), ஐரோப்பாவில் பிற (எ.கா. சார்ப்ரூக்கன் மற்றும் ஸ்டட்கர்ட் (Saarbrücken and Stuttgart) மற்றும் வட அமெரிக்காவில் (North America) பல. மிக முக்கியமான குழு பிட்ஸ்பர்க்கில் (Pittsburgh) உள்ள கார்னகி-மெலன் பல்கலைக்கழகத்தில் (Carnegie-Mellon University) ஜெய்ம் கார்பனெல் மற்றும் செர்ஜி நிரன்பேர்க்கின் (Jaime Carbonell and Sergei Nirenburg) கீழ் இருந்தது; பல அறிவு சார்ந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளுடன் (குட்மேன் மற்றும் நிரன்பர்க் (Goodman and Nirenburg) 1991) பரிசோதனை செய்யப்பட்டது. அடிப்படை ஒழுங்குமுறை கூறுகள் (basic system components) பொருண்மைக்களத்திற்கான (domain) ஒரு சிறிய கருத்து அகராதி, இரண்டு மொழிகளுக்கும் ஒரு பகுப்பாய்வு மற்றும் ஒரு உருவாக்க அகராதி, பொருண்மையியல்சார் கட்டுப்பாடுகள் (semantic constraints) கொண்ட ஒரு தொடரியல் பாகுபடுத்தி (syntactic parser), ஒரு பொருண்மையியல்சார் பொருத்தி (semantic mapper) (பொருண்மையியல்சார்

பொருள்கோளுக்கு), ஒரு ஊடாடும் 'அதிகப்படுத்தி (interactive augmentor)', சொல்சார் தேர்வுடன் (lexical selection) இலக்குமொழி தொடரியல் கட்டமைப்புகளை உருவாக்கும் பொருண்மையியல்சார் உருவாக்கி (semantic generator) மற்றும் இலக்குமொழி வாக்கியங்களை உருவாக்குவதற்கான தொடரியல்சார் உருவாக்கி (syntactic generator) என்பனவாகும். கருத்து அகராதி மற்றும் பகுப்பாய்வு மற்றும் உருவாக்க அகராதிகளில் (அதாவது பொருண்மையியல்சார் கட்டுப்பாடுகளை வரையறை விளக்கம்செய்தல்) உள்ள பொருண்மையியல்சார் தகவல்கள் மொழி-சுதந்திரமானவை ஆனால் ஒரு பொருண்மைக்களத்திற்கு என்று குறிப்பிடப்பட்டது. ஒழுங்குமுறையின் மையம், பொருண்மைக்கள 'கருத்து அகராதி' அறிவுக் குறிப்புடன் 'அதிகப்படுத்தியால்' நிகழ்த்தப்பட்ட பொருண்மையியல்சார் பகுப்பாய்வு மற்றும் ஊடாடும் பொருண்மை மயக்கநீக்கம் (interactive disambiguation) இவற்றின் செயல்முறைகளிலிருந்து ஆக்கப்பட்ட கருத்துரைகளின் வலைப்பின்னல்களின் (networks of propositions) வடிவத்தில் இருக்கின்ற பனுவல்களின் இடைமொழிய உருப்படுத்தம் ஆகும். 1980களின் முடிவில், கார்னகி-மெலன் (Carnegie-Mellon) அணி முழுமையாக அதன் KANT முன்மாதிரி ஒழுங்குமுறையை விரிவுபடுத்தியது மற்றும் கேட்டர்பில்லர் கார்பரேஷனுக்காக (Caterpillar Corporation) ஒரு செயல்பாட்டு அறிவு அடிப்படையிலான ஒழுங்குமுறையின் (operational knowledge-based system) உருவாக்கத்தைத் தொடங்கத் தயாராக இருந்தது - ஒட்டுமொத்த தரத்தை மேம்படுத்துவதற்காக ஒரு நிறுவனம் உருவாக்கிய கட்டுப்பாட்டு மொழியை (company-developed controlled language) உட்படுத்தியது.

1980களின் நடுப்பகுதியில் இருந்து 'ஒன்றிணைத்தல்' ('unification') மற்றும் 'கட்டுப்பாடு'சார் ('constraintbased') வடிவவதங்களை ஏற்றுக்கொள்ளும் ஒரு போக்கு இருந்தது [எ.கா. சொல்சார்-செயல்பாட்டு இலக்கணம் (Lexical-Functional Grammar), தலை-இயக்க தொடரமைப்பு இலக்கணம் (Head-Driven Phrase Structure Grammar), வகைபாட்டு இலக்கணம் (Categorial Grammar) முதலியன]. சிக்கலான பன்னிலை உருப்படுத்தங்களுக்கும் (multi-level representation) மாற்றங்களின் பெரிய கணங்களுக்கும் (large sets of transformation) அருவ விதிகளின் கட்டுப்படுத்தப்பட்ட கணங்களுக்கும் (restricted set of abstract rules) பதிலாக குறிப்பிட்ட சொற்களில் உட்படுத்தப்பட நிபந்தனைகள் மற்றும் கட்டுப்பாடுளுடன் ஒற்றை

அடுக்குசார் உருப்படுத்தங்கள் (mono-stratal representations) இருந்தன. இது பகுப்பாய்வு, மாற்றம் மற்றும் உருவாக்கம் இவற்றை எளிமைப்படுத்த வழிவகுத்தது; அதே நேரத்தில் இந்த இலக்கணங்களின் கூறுகள் கொள்கை அடிப்படையில் உள்ளன தலைகீழ் மாற்றத்திகுரியன. கடந்த காலங்களில் பரிமாற்ற ஒழுங்குமுறைகளைப் (transfer systems) பண்பாகம் செய்த தொடரியல் நோக்குநிலை மாற்றப்பட்டது சொல்லியலார் ('லெக்சிகலிஸ்ட்' ('lexicalist') அணுகுமுறைகளால் இடம்பெயர்க்கப்பட்டது; இதன் விளைவாக சொல்சார் அலகுகளுடன் இணைக்கப்பட்ட தகவல்களின் வரம்பில் அதிகரிப்பு ஏற்பட்டது: உருபனியல் மற்றும் இலக்கணம்சார் தரவு மற்றும் மொழிபெயர்ப்பு நிகரன்கள் மட்டுமல்லாமல், தொடரியல்சார் மற்றும் பொண்மையியல்சார் கட்டுப்பாடுகள் பற்றிய தகவல்கள் மற்றும் மொழியியல்சாரா மற்றும் கருத்துருசார் தகவல்கள் (conceptual information).

சொல்சார் தரவின் விரிவாக்கம் இடைமொழி அடிப்படையிலான ஒழுங்குமுறைகளின் அகராதிகளில் மிகத் தெளிவாகக் காணப்படுகிறது; இது மொழியியல் அல்லாத பெரிய அளவிலான தகவல்களை அடக்கும். பல குழுக்கள் மொழி கற்பவர்களுக்கான இருமொழி அகராதிகள், பொது ஒருமொழிய அகராதிகள், சிறப்பு தொழில்நுட்ப அகராதிகள் (specialized technical dictionaries) மற்றும் தொழில்முறை மொழிபெயர்ப்பாளர்களால் பயன்படுத்தப்படும் கலைச்சொல் தரவுவங்கிகள் (terminological databanks) போன்ற எளிதில் கிடைக்கக்கூடிய அகராதிசார் மூலங்களிலிருந்து சொல்சார் தகவல்களைப் பிரித்தெடுக்கும் முறைகளில் ஆய்வதையும் மற்றும் ஒத்துழைப்பையும் செய்தது. இந்தப் புலத்தில் ஒரு குறிப்பிடத்தக்க முயற்சி 1980களின் பிற்பகுதியில் பல ஜப்பானிய கணினி உற்பத்தி நிறுவனங்களால் ஆதரிக்கப்பட்ட மின்னணு அகராதி ஆராய்ச்சி திட்டம் (Electronic Dictionary Research project) ஆகும், இந்தச் சொல்சார் செயல்பாடு தற்போதைய காலத்திற்கும் தொடர்கிறது. பெரிய தரவுத்தள மற்றும் அகராதி மூலவளங்கள், மொழியியல் தரவு கூட்டமைப்பு (Linguistic Data Consortium) (யுஎஸ்ஏயில்) மற்றும் ஐரோப்பிய மொழி மூலவளங்கள் சங்கம் (European Language Resources Association (ELRA)) என்பன மூலம் கிடைக்கின்றன; இந்த நிறுவனம் மொழி மூலவளங்கள் மற்றும் மதிப்பீட்டு மாநாடுகள் (Language Resources and Evaluation Conferences (LREC))

என்ற தலைப்புக்கு அர்ப்பணிக்கப்பட்ட ஒரு பெரிய இருபதாண்டு தொடர் மாநாடுகளைத் தொடங்கி வைத்தது.

2.8. மொழிபெயர்ப்பு கருவிகள் மற்றும் மொழிபெயர்ப்பாளரின் பணிநிலையம்

ஹட்சின்ஸ் (Hutchins 2015) "மொழிபெயர்ப்பு கருவிகள் மற்றும் மொழிபெயர்ப்பாளரின் பணிநிலையம் (Translation tools and the translator's workstation)" என்ற தலைப்பின் கீழ் தந்துள்ள செய்திகள் இங்கு தொகுத்துத் தரப்பட்டுள்ளன. 1980களில் மொழிபெயர்ப்பாளர்கள் தங்கள் பணிக்கான கணினிகளின் நன்மைகளைப் பற்றி அறிந்திருந்தனர் - சொல் செயலாக்கம் (word processing), தனிப்பட்ட பொருள்விளக்கச் சொற்கோவைகள் (glossaries) உருவாக்குதல், நிகழ்நிலை on-line access) அணுகலுக்கான மற்றும் ஆவணங்களின் பரப்புகைக்கான (transmission) வசதிகள். எவ்வாறாயினும், அவர்கள் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் வெளியீட்டின் தரத்தில் திருப்தி அடையவில்லை. மொழிபெயர்ப்பாளர்கள் கணினி உதவிக்கருவியை அவர்கள் செயல்முறைகளின் கட்டுப்பாட்டில் வைத்திருக்க விரும்புகிறார்கள் என்றும் தானியங்கி அமைப்புகளின் 'அடிமைகளாக' இருக்க விரும்பவில்லை என்றும் ஏற்கனவே தெளிவாகத் தெரிந்தது. பல கருவிகள் உருவாக்கப்பட்டன, குறிப்பாக சொல்லடைவு ஆக்கம் (concordancing), அகராதி உருவாக்கம், கலைச்சொல் மேலாண்மை (terminology management) மற்றும் ஆவணப் பரப்புகை ('translation memories'). இருப்பினும் 1990களின் முற்பகுதியில் வந்த மிக முக்கியமான வளர்ச்சி, 'மொழிபெயர்ப்பாளரின் பணிநிலையத்தில்' ('translator's workstation') (அல்லது 'வொர்க் பெஞ்ச்'/பணிமேடை ('workbench')) ஒருங்கிணைந்த கருவிகளின் சந்தைப்படுத்தல்.

மொழிபெயர்ப்பு பணிநிலையங்கள் (translation workstations) பன்மொழி சொல் செயலாக்கம் (multilingual word processing), ஒளி எழுத்துணரி (Optical Character Recognition (OCR)) வசதிகள், கலைச்சொல் மேலாண்மை மென்பொருள் (terminology management software), சொல்லடைவு ஆக்கத்திற்கான (concordancing) வசதிகள் மற்றும் குறிப்பாக 'மொழிபெயர்ப்பு நினைவுகள்' ('translation memories') ஆகியவற்றை இணைக்கின்றன. பிந்தைய வசதி மொழிபெயர்ப்பாளர்களுக்கு அசல் பனுவல்களையும் அவற்றின் மொழிபெயர்க்கப்பட்ட

பதிப்புகளையும் அருகருகே சேமிக்க உதவுகிறது, அதாவது மூல மற்றும் இலக்கின் தொடர்புடைய வாக்கியங்கள் வரிசைப்படுத்தப்படுகின்றன. மொழிபெயர்ப்பாளர் இவ்வாறு மொழிபெயர்ப்பு நினைவகத்தில் ஒரு மொழியின் சொற்றொடர்களையோ அல்லது முழு வாக்கியங்களையோ தேடலாம் மற்றும் பிற மொழியில் அதனுடன் தொடர்புடைய சொற்றொடர்களை, சரியான பொருத்தங்களை அல்லது தோராயங்களைக் காட்டலாம். கூடுதலாக, மொழிபெயர்ப்பு பணிநிலையங்கள் பெரும்பாலும் (கூறுகள், பத்திகள் அல்லது முழு பனுவல்களையும் மொழிபெயர்க்க) மொழிபெயர்ப்பாளர்கள் பொருத்தமானவைகளாகப் பயன்படுத்துகின்ற அல்லது தெரிந்தெடுக்கின்ற முழு இயந்திர மொழிபெயர்ப்பு நிரல்களைத் தருகின்றன. மொழிபெயர்ப்பாளர்களுக்கான கணினி அடிப்படையிலான இந்த பல்வேறு வசதிகளின் அடிப்படை யோசனைகள் 1980களின் முற்பகுதிக்குச் செல்கின்றன (Hutchins 1998ஐப் பார்க்கவும்).

பணிநிலையங்களில் இப்போது பல விற்பனையாளர்கள் உள்ளனர். முந்தையவை டிராடோஸ் (Trados (மொழிபெயர்ப்பாளர்களின் பணிமேடை (Translator's Workbench)), ஸ்டார் ஏஜி (STAR AG) (Transit), ஐபிஎம் (IBM) (மொழிபெயர்ப்பு மேலாளர் (TranslationManager), சந்தைப்படுத்தப்படவில்லை), யூரோலாங் ஆப்டிமைசர் (யுரோலாங் மேம்படுத்தி (Eurolang Optimizer)) (கிடைப்பதில்லை) என்பன முந்தவைகள் ஆகும். 1990களில் மற்றும் 2000களின் முற்பகுதியில் இன்னும் பல தோன்றின: அட்ரில் (Atril) (Déjà Vu/ டிஜோ வு), எஸ்.டி.எல். (SDL) (எஸ்.டி.எல்.எக்ஸ் ஒழுங்குமுறை (SDLX system)), ஜெராக்ஸ் (Xerox) (எக்ஸ்.எம்.எஸ் (XMS)), டெர்மினோடிக்ஸ் (Terminotix) (லாஜிடெர்ம் (LogiTerm)), மல்டிகார்போரா (MultiCorpora) (மல்டிட்ரான்ஸ் (MultiTrans), சாம்போலியன் (Champollion) (வேர்ட்ஃபாஸ்ட் (WordFast)), மெட்டாடெக்ஸிஸ் (MetaTexis) மற்றும் புரோமெமோரியா (ProMemoria). மொழிபெயர்ப்பு பணிநிலையம் மொழிபெயர்ப்பாளர்களின் கணினிகள் பயன்பாட்டில் புரட்சியை ஏற்படுத்தியுள்ளது; அவர்கள் தேர்வு செய்யும் எந்தவொரு (அல்லது எதுவுமில்லை) வசதியையும் பயன்படுத்தி தங்கள் முழு கட்டுப்பாட்டில் இருக்கும் ஒரு கருவியை இப்போது வைத்திருக்கிறார்கள்.

2.9. 1989 முதல் ஆராய்ச்சி

ஹட்சின்ஸ் (Hutchins 2015) ”1989 முதல் ஆராய்ச்சி (Research since 1989)” என்ற தலைப்பின் கீழ் கூறியுள்ள செய்திகள் இங்கு தொகுத்துத் தரப்பட்டுள்ளன. 1980களின் இறுதி வரை இயந்திரமொழிபெயர்ப்பு ஆராய்ச்சியின் மேலாதிக்க சட்டகமானது பல்வேறு வகையான மொழியியல் விதிகள் அடிப்படையில் அமைந்தது: தொடரியல் பகுப்பாய்விற்கான விதிகள் (rules for syntactic analysis), சொல்சார் இடமாற்றத்திற்கான விதிகள் (rules for lexical transfer), தொடரியல்சார் உருவாக்கத்திற்கான விதிகள் (rules for syntactic generation), உருபனியல் விதிகள் (rules for morphology), சொல்சார் விதிகள் (lexical rules) போன்றவை. விதி அடிப்படையிலான அணுகுமுறை அரியேன் (Ariane), மெட்டல் (METAL), சூசி (SUSY), மு (Mu) மற்றும் யூரோட்ரா (Eurotra) இவற்றின் இடமாற்ற ஒழுங்குமுறைகளில் (transfer systems) மிகவும் தெளிவாக இருந்தது; ஆனால் இது பல்வேறு இடைமொழி ஒழுங்குமுறைகள் (interlingua systems) அடிப்படையிலும் இருந்தது. இவை இரண்டும் அடிப்படையில் மொழியியல் சார்ந்தவை [(டி.எல்.டி (DLT) மற்றும் ரொசெட்டா (Rosetta)] மற்றும் அறிவு அடிப்படையிலானவை (KANT). இருப்பினும், 1989முதல் விதி அடிப்படையிலான அணுகுமுறையின் ஆதிக்கம் புதிய முறைகள் மற்றும் உத்திகள் தோன்றியதன் மூலம் உடைக்கப்பட்டுள்ளன, அவை இப்போது ‘தரவுத்தொகுதி அடிப்படையிலான’ நெறிமுறைகள் (‘corpusbased’ methods) என்று அழைக்கப்படுகின்றன.

2.9.1. தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள் (Corpus-based approaches)

மிகவும் வியத்தகு வளர்ச்சியானது புள்ளியியல் அடிப்படையிலான அணுகுமுறைகளின் (statistics-based approaches) புத்துயிர் ஆகும் - இது முதல் பதின்ம ஆண்டின் (மேலே உள்ள பிரிவு 2.3) ‘அனுபவவாதத்திற்கு’ (‘empiricism’) திருப்பிவரவாகக் கருதப்படுகிறது மற்றும் முன்னர் 1970கள் மற்றும் 1980களில் ஆதிக்கம் செலுத்திய விதி அடிப்படையிலான ‘பகுத்தறிவு வாதத்திற்கு’ (‘rationalism’) ஒரு சவாலாக அமைகின்றது. பேச்சு அறிதலில் வாய்ப்பியல்சார் நுட்பங்களின் (stochastic techniques) வெற்றியுடன், ஐபிஎம்-இல் [யார்க்க்டவுன் ஹைட்ஸ், நியூயார்க் (Yorktown Heights, New York)] உள்ள ஒரு குழு இயந்திர மொழிபெயர்ப்பில் அவைகளின் பயன்பாட்டை மீண்டும் பார்க்கத் தொடங்கியது. அவர்களின் கேண்டைட் ஒழுங்குமுறையின் (Candide system) தனித்துவமான அம்சம் என்னவென்றால், புள்ளியியல்சார் நெறிமுறைகள் (statistical methods) கிட்டத்தட்ட பகுப்பாய்வு மற்றும் உருவாக்கத்தின் ஒரே வழிமுறையாகப் பயன்படுத்தப்பட்டன; மொழியியல் விதிகள் எதுவும் பயன்படுத்தப்படவில்லை. ஐபிஎம் ஆராய்ச்சி அடிப்படையாகக் கொண்டது கனேடிய பாராளுமன்ற விவாதங்களின் அறிக்கைகளில் (reports of Canadian parliamentary debates) [தி கனடிய ஹான்சார்ட் (the Canadian Hansard)] உள்ள பிரெஞ்சு மற்றும் ஆங்கில பனுவல்களின் தரவுத்தொகுதி

அடிப்படையிலானதாகும். இந்த முறையின் சாராம்சம் முதலில் இணையான பனுவல்களின் (parallel texts) தொட்டர்கள், சொல் குழுக்கள் மற்றும் தனிச் சொற்கள் இவற்றை வரிசைப்படுத்துவதாகும்; பின்னர் ஒரு மொழியின் வாக்கியத்தின் ஒரு சொல் அது வரிசைப்படுத்ததப்பட்ட மொழியுடன் (ஒரு 'மொழிபெயர்ப்பு மாதிரி') மொழிபெயர்க்கப்பட்ட வாக்கியத்தில் உள்ள ஒரு சொல் அல்லது சொற்களுடன் ஒத்திருக்கிறது என்பதன் நிகழ்தகவுகளைக் (probabilities) கணக்கிடுவதாகும். வெளியீடுகள் பின்னர் சரிபார்க்கப்பட்டு இலக்குமொழியில் இருமொழி பனுவல்களின் தரவுத்தொகுதியிலிருந்து (corpus of bilingual texts) [ஒரு 'மொழி மாதிரி' ('language model')] ஆக்கப்பட்ட சொல்லிலிருந்துசொல் நிலைமாற்ற நிகழ்வெண் (word-to-word transition frequencies) அடிப்படையில் மறுசீரமைக்கப்பட்டன. மொழியியல் அடிப்படையிலான முறைகளில் வளர்ந்த பெரும்பாலான ஆராய்ச்சியாளர்களுக்கு கிடைத்த ஆச்சரியம் என்னவென்றால், முடிவுகள் மிகவும் ஏற்றுக்கொள்ளத்தக்கவையாக இருந்தன: மொழிபெயர்க்கப்பட்ட கிட்டத்தட்ட பாதி சொற்றொடர்கள் தரவுத்தொகுதியில் உள்ள மொழிபெயர்ப்புகளுடன் சரியாக பொருந்தின, அல்லது அதே உணர்வை சற்று வித்தியாசமான வார்த்தைகளில் வெளிப்படுத்தியது அல்லது பிற சமமான முறையான மொழிபெயர்ப்புகளை வழங்கியது.

இக்காலத்திலிருந்து புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு (statistical machine translation (SMT)) பலரின் முக்கிய கவனக்குவிப்பாக மாறியது; ஆராய்ச்சி குழுக்கள், முதன்மையாக ஐபிஎம் மாதிரியை அடிப்படையாகக் கொண்டன; ஆனால் பல அடுத்தடுத்த சீரமைப்புகளைக் கொண்டிருந்தன. அசல் மூல மற்றும் இலக்கு மொழிகளுக்கு இடையிலான சொல் பொருத்தங்களின் (correlations) முக்கியத்துவம் 'தொடர்களுக்கு' இடையிலான (அதாவது சொற்களின் வரிசைமுறைகள், 'பாரம்பரிய' பெயத்தொடராக இருக்க வேண்டியதில்லை, வினைத் தொடர்கள் அல்லது முன்னுருபுத் தொடர்கள்) பொருத்தங்களால் (correlations) இடம்பெயர்க்கப்பட்டது; உருபனியல் மற்றும் தொடரியல் தகவல்களைச் சேர்ப்பதன் மூலமும் பயன்படுத்துவதன் மூலமும் அகராதி மற்றும் சொற்களஞ்சிய (thesaurus) மூலவளங்கள் மூலமும் மேம்படுத்தப்பட்டன. மொழி மாதிரிகளில் பயன்படுத்தப்படும் வரிசைப்படுத்தப்பட்ட இருமொழிய தரவுத்தளங்களின் மற்றும் ஒருமொழியத் தரவுத்தொகுதியின் அளவுகளில் பரந்த அதிகரிப்பு ஏற்பட்டுள்ளது; எஸ்எம்டி அணுகுமுறை (SMT approach) மொழி இணைகளின் பரந்துவரும் பரப்பெல்லையில் பயன்படுத்தப்படுகிறது. எஸ்எம்டி ஆராய்ச்சிக்கான (SMT research) முக்கிய மையங்கள் ஆச்சென் மற்றும் தெற்கு கலிபோர்னியா பல்கலைக்கழகங்களாகும் (universities of

Aachen and Southern California); சாமீபத்தில் அவற்றுடன் கூகிள் கார்ப்பரேஷன் (Google Corporation) இணைந்துள்ளது.

பனுவல் தரவுத்தொகுதிகளின் பெரிய தரவுவங்கிகளுக்கு மேம்பட்ட விரைவான அணுகலிலிருந்து பயனடைகிற இரண்டாவது பெரிய 'தரவுத்தொகுதி அடிப்படையிலான' அணுகுமுறை 'எடுத்துக்காட்டு அடிப்படையிலான' ('example-based') (அல்லது 'நினைவக அடிப்படையிலான' ('memory-based')) என அழைக்கப்படுகிறது. 1981இல் மாகோடோ நாகோவால் (Makoto Nagao) முதன்முதலில் முன்மொழியப்பட்டாலும் 1980களின் இறுதியில் மட்டுமே அதன் சோதனைகள் தொடங்கப்பட்டன; ஆரம்பத்தில் சில ஜப்பானிய குழுக்களிலும் டி.எல்.டி திட்டத்திலும் (DLT project) (மேலே உள்ள பிரிவு 2.7). எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் (example-based machine translation (EBMT)) அடிப்படைக் கருதுகோள், மொழிபெயர்ப்பு பெரும்பாலும் ஒத்த உதாரணங்களைக் கண்டறிதல் அல்லது நினைவுபடுத்துதல் ஆகியவற்றை ஈடுபடுத்தும்; அதாவது ஒரு குறிப்பிட்ட வெளிப்பாடு அல்லது சில ஒத்த நிகரான தொடர்கள் அல்லது சொல் குழுக்கள் எவ்வாறு இதற்கு முன் மொழிபெயர்க்கப்பட்டுள்ளது என்பதாகும். அணுகுமுறை, புள்ளியியல்சார் முறைகளால் (SMTஇல் பயன்படுத்தப்பட்டதைப் போன்றது) அல்லது பாரம்பரிய விதி அடிப்படையில் வரிசைப்படுத்தப்பட்ட இணையான இருமொழி பனுவல்களின் தரவுத்தளத்திலிருந்து சமமான சொற்றொடர்கள் அல்லது சொற்குழுமங்களைப் பிரித்தெடுக்கும் மற்றும் தேர்ந்தெடுக்கும் செயல்முறைகளில் நிறுவப்பட்டுள்ளது. பொருத்தங்களைக் கணக்கிடுவதற்குச் சில குழுக்கள் பொருண்மையியல்சார் முறைகளைப் பயன்படுத்துகின்றன; எ.கா. ஒரு பொருண்மையியல்சார் வலைப்பின்னல் (semantic network) அல்லது ஒரு பொருண்மைக்களச் சொற்களின் படிநிலை (சொற்களஞ்சியம்) *hierarchy (thesaurus) of domain terms); பிற குழுக்கள் இலக்கு மொழியில் சொல்சார் நிகழ்வெண்களைப் பற்றிய புள்ளியியல்சார் தகவல்களைப் பயன்படுத்துகின்றன. சரளமாகவும் இலக்கணம்சார் வெளியீட்டையும் உருவாக்குவதற்காக தேர்ந்தெடுக்கப்பட்ட இலக்கு மொழி எடுத்துக்காட்டுகளின் (பொதுவாக குறுகிய தொடர்கள்) மறு சேர்கைதான் (re-combination) ஒரு பெரிய சிக்கல் ஆகும். ஆயினும்கூட அணுகுமுறையின் முக்கியமான நன்மை (விதி அடிப்படையிலான அணுகுமுறைகளுடன் ஒப்பிடுகையில்) பனுவல்கள் தொழில்முறை

மொழிபெயர்ப்பாளர்களால் (professional translators) உருவாக்கப்பட்ட உண்மையான மொழிபெயர்ப்புகளின் தரவுத்தளங்களிலிருந்து பிரித்தெடுக்கப்படுவதால் விளைவுகள் மரபுத்தன்மையாக இருக்கும். எஸ்எம்டி (SMT) போலல்லாமல் ஒரு 'வழக்கமான' இபிஎம்டி (EBMT) மாதிரி எதுவாக இருக்கும் என்பதில் சிறிய உடன்பாடு இல்லை மற்றும் எடுத்துக்காட்டு அடிப்படையிலான முறைகளுக்கு அர்ப்பணிக்கப்பட்ட பெரும்பாலான ஆராய்ச்சி எந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையிலும் பயன்படுத்தக்கூடியது.

இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சிக்கான எஸ்எம்டி (SMT) அப்போது ஆதிக்கம் செலுத்தும் கட்டமைப்பாக இருந்தாலும் தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள் இரண்டும் பல விஷயங்களில் ஒன்றிணைந்து வருகின்றன; ஏனெனில் எஸ்எம்டி தொடர் அடிப்படையிலான வரிசைப்படுத்தலையும் மொழியியல் தரவையும் அதிகம் பயன்படுத்துகிறது; ஈபிஎம்டி புள்ளிவிவர நுட்பங்களை பரவலாகப் பயன்படுத்துகிறது என இவை அங்கீகரிக்கப்பட்டுள்ளன. இதன் விளைவாக இரண்டு மாதிரிகளின் தனித்துவமான அம்சங்களை தனிமைப்படுத்துவது மிகவும் கடினமாகி வருகிறது.

இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சிக்கான தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள் பெரும் அளவில் இருமொழி மற்றும் பன்மொழி பனுவல் தரவுத்தொகுதி (உண்மையில் ஒருமொழி தரவுத்தொகுதி) கிடைப்பதை நம்பியுள்ளதால் கடந்த பதின்ம ஆண்டில் அல்லது அதற்கு மேற்பட்டவற்றில் பனுவல் தரவுத்தளங்களின் சேகரிப்பு மற்றும் மதிப்பீட்டில் கவனம் செலுத்தப்பட்டன. மற்றும் புள்ளியியல்சார் தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள் கணினி மொழியியல் மற்றும் இயற்கையான மொழி ஆய்வு இவற்றின் பரந்த துறையில் ஆதிக்கம் உள்ளதாக மாறின. மொழியியல் மூலவளங்கள் அப்போது இயந்திர மொழிபெயர்ப்பு மற்றும் இயற்கையான மொழி ஆய்வு இரண்டிற்கும் மையமாக இருந்தன; இதன் விளைவாக, இயந்திர மொழிபெயர்ப்பு அப்போது கணினி மொழியியலின் 'பிரதான நீரோட்டத்திற்கு' ('mainstream') திரும்பியது - இந்த நிலையை ALPAC அறிக்கையின் பின்னர் இழந்தது (மேலே உள்ள பிரிவு 2.4 ஐப் பார்க்கவும்); மேலும் இது ஏற்கனவே குறிப்பிடப்பட்ட மொழியியல் தரவு கூட்டமைப்பு (Linguistic Data Consortium), ஐரோப்பிய மொழி முலவளங்கள் சங்கம் (European Language Resources Association (LREC)), இருபதாண்டு காலிங் மாநாடுகள்

(biennial Coling conferences), மற்றும் இருபதாண்டு மொழி முலவளங்கள் மற்றும் மதிப்பீட்டு மாநாடுகள் (biennial Language Resources and Evaluation Conferences (LREC)) ஆகியவற்றின் செயல் எல்லையில் பிரதிபலித்தது.

2.9.2. விதி அடிப்படையிலான அணுகுமுறைகள்

1990 முதல் முக்கிய கண்டுபிடிப்பு தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகளின் வளர்ச்சியாக இருந்தாலும் விதி அடிப்படையிலான ஆராய்ச்சி இடமாற்றம் மற்றும் இடைமொழிய ஒழுங்குமுறைகள் இரண்டிலும் தொடர்ந்தது. யூரோட்ராவில் ஈடுபட்ட பல ஆராய்ச்சியாளர்கள் வளர்ந்த கோட்பாட்டு அணுகுமுறையில் செயல்பட்டனர்; எ.கா. சார்ப்ரூக்கனில் (Saarbrücken) காட்2 ஒழுங்குமுறை (CAT2 system); யூரோட்ரா ஆராய்ச்சியின் பலன்களில் ஒன்று டேனிஷ்/ஆங்கில மொழிபெயர்ப்பு காப்புரிமைக்கு வேண்டி டென்மார்க்கில் உருவாக்கப்பட்ட பாட்ரான்ஸ் இடமாற்ற அடிப்படையிலான ஒழுங்குமுறை (PaTrans transfer-based system) ஆகும்.

மொழியியல் அடிப்படையிலான இடமாற்ற அணுகுமுறையின் (linguistics-based transfer approach) மற்றொரு எடுத்துக்காட்டு எல்எம்டி (LMT) திட்டம் ஆகும்; இது 1980களின் நடுப்பகுதியில் மைக்கேல் மெக்கார்டின் (Michael McCord) கீழ் ஜெர்மனி, ஸ்பெயின், இஸ்ரேல் மற்றும் அமெரிக்கா இவற்றில் உள்ள பல ஐபிஎம் ஆராய்ச்சி மையங்களில் தொடங்கப்பட்டது. புரோலாகில் (Prolog) செயல்படுத்தப்பட்ட எல்எம்டி-க்கு [தருக்க நிரலாக்க இயந்திர மொழிபெயர்ப்பு (LMT ('Logic-programming Machine Translation')] பாரம்பரிய நான்கு நிலைகள் உள்ளன: சொல்சார் பகுப்பாய்வு (lexical analysis); புற மற்றும் ஆழமான (தருக்க) உறவுகளின் உருப்படுத்தங்களை உருவாக்கும் மூலப் பனுவல்களின் தொடரியல் பகுப்பாய்வு (syntactic analysis); சம உருவுள்ள கட்டமைப்புப் இடமாற்றம் மற்றும் மாற்றங்களை மறு அமைப்பாக்கம் செய்யும் இடமாற்றம் (transfer); மற்றும் இலக்குப் பனுவல்களின் உருபனியல் உருவாக்கம் (generation).

கார்னகி-மெலன் பல்கலைக்கழகத்தில் (Carnegie-Mellon University (CMU)) இடைமொழிய அணுகுமுறை தொடர்ந்தது. 1992இல் ஒரு பெரிய அளவிலான அறிவு

அடிப்படையிலான மற்றும் கட்டுப்படுத்தப்பட்ட மொழி தொழில்நுட்ப கையேடுகளின் பன்மொழிய மொழிபெயர்ப்பிற்கான கேடலிஸ்ட் (CATALYST) ஒழுங்குமுறைக்காக கேட்டர்பில்லர் நிறுவனத்துடன் (Caterpillar company) ஒத்துழைப்பைத் தொடங்கியது. பதின்ம ஆண்டு இறுதியில் CMUஇல் அறிவு அடிப்படையிலான அணுகுமுறை விரைவான முன்மாதிரிக்கும் சிறப்பு-நோக்க ஒழுங்குமுறைகளின் (டிப்ளோமாட் (DIPLOMAT)) செயல்படுத்தலுக்கும் வேண்டி பனுவல் தரவுத்தொகுதியின் புள்ளியியல்சார் பகுப்பாய்வின் முன்னேற்றங்களுடன் இணைக்கப்பட்டது, எ.கா. இராணுவ நடவடிக்கைகளில் செர்போ-குரோஷியனின் மொழிபெயர்ப்பு (translation of Serbo-Croatian in military operations).

1990களின் நடுப்பகுதியில் பிற இடைமொழிய அடிப்படையிலான ஒழுங்குமுறைகள் தொடங்கப்பட்டன, எ.கா. புதிய மெக்ஸிகோ மாநில பல்கலைக்கழகத்தில் (New Mexico State University) செர்ஜி நிரன்பேர்க்கால் (Sergei Nirenburg) உருவாக்கப்பட்ட அல்ட்ரா (ULTRA) ஒழுங்குமுறை, கொள்கைகள் மற்றும் அளவுருக்களின் மொழியியல் கோட்பாடு (linguistic theory of Principles and Parameter) அடிப்படையில் மேரிலாந்து பல்கலைக்கழகத்தில் (University of Maryland) போனி ஜே. டோர்-ஆல் (Bonnie J. Dorr) (1993)உருவாக்கப்பட்ட யுனிட்ரான் ஒழுங்குமுறை (UNITRAN system) மற்றும் தெற்கு கலிபோர்னியா (Southern California), நியூ மெக்ஸிகோ மாநிலம் (New Mexico State) மற்றும் கார்னகி-மெலன் (Carnegie-Mellon) பல்கலைக்கழகங்களை உள்ளடக்கிய கூட்டு திட்டமான பாங்லோஸ் திட்டம் (Pangloss project).

பாங்லோஸ் (Panglos), அர்பாவால் [மேம்பட்ட ஆராய்ச்சி திட்டங்கள் நிறுவனம் (Advanced Research Projects Agency (ARPA)] ஆதரிக்கப்பட்ட மூன்று இயந்திர மொழிபெயர்ப்புத் திட்டங்களில் ஒன்றாகும்; மற்றவை மேலே குறிப்பிட்டுள்ள ஐபிஎம் புள்ளியியல் அடிப்படையிலான திட்டம் (IBM statistics-based project) மற்றும் டிராகன் சிஸ்டம்ஸ் (Dragon Systems) உருவாக்கிய ஒழுங்குமுறை. இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சிக்கு அமெரிக்க அரசாங்கத்தின் ஆதரவை மறுசீரமைப்பது ALPAC அறிக்கையின் சேதப்படுத்தும் தாக்கத்தின் முடிவைக் குறிக்கிறது (மேலே உள்ள பிரிவு 2.5).

இறுதியாக, 1990களின் இறுதியில், ஐக்கிய நாடுகளின் மேம்பட்ட ஆய்வுகள் நிறுவனம் பல்கலைக்கழகம் (Institute of Advanced Studies of the United Nations University) (டோக்கியோ (Tokyo)) அதன் பன்னாட்டு இடைமொழிய அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு திட்டத்தை தொடங்கியது (multinational interlingua based MT project) - இது ஒரு 'தரப்படுத்தப்பட்ட' இடைநிலை மொழி யுஎன்எல் [உலகளாவிய வலைப்பின்னல் மொழி (Universal Networking Language (UNL)] அடிப்படையில் அமையும்; ஆரம்பத்தில் ஐக்கிய நாடுகள் சபையின் ஆறு அதிகாரப்பூர்வ மொழிகளுக்கும் ஆறு பரவலாக பேசப்படும் மொழிகளுக்கும் (அரபு, சீன, ஆங்கிலம், பிரஞ்சு, ஜெர்மன், இந்தி, இந்தோனேசிய, இத்தாலியன், ஜப்பானிய, போர்த்துகீசியம், ரஷ்ய மற்றும் ஸ்பானிஷ்) தொடங்கப்பட்டது - சில 15 நாடுகளின் குழுக்களை உள்ளடக்கியது.

2.9.3. பேச்சு மொழிபெயர்ப்பு

1980களின் பிற்பகுதியிலிருந்து மிக முக்கியமான முன்னேற்றங்களில் ஒன்று பேச்சு மொழி மொழிபெயர்ப்பில் (spoken language translation) வளர்ந்து வரும் ஆர்வம் ஆகும்; இது பேச்சு புரிந்துகொள்ளலையும் (speech recognition) பேச்சு உருவாக்கத்தையும் (speech synthesis) இணைப்பது, கலந்துரையாடல்கள் மற்றும் உரையாடல்களின் பொருள்கோள், பொருண்மையியல் பகுப்பாய்வு மற்றும் சமூக சூழல்கள் மற்றும் சூழ்நிலைகளுக்கு உணர்திறன் இவற்றின் வல்லமையான சவால்களை முன்வைக்கிறது. 1980களின் பிற்பகுதியில் பிரிட்டிஷ் டெலிகாம் (British Telecom) பேச்சு மொழி தொடர்நூல் வகை ஒழுங்குமுறையில் (spoken language phrasebook type system) சில சோதனைகளை செய்தது. இருப்பினும் முதல் நீண்டகால குழு 1986இல் ஏ.டி.ஆரில் நிறுவப்பட்டது [(தொலைத்தொடர்பு ஆராய்ச்சி ஆய்வகங்களை விளக்குதல் (Interpreting Telecommunications Research Laboratories)] [ஜப்பானில் (Japan) ஒசாகாவிற்கு (Osaka) அருகிலுள்ள நாராவை (Nara) சார்ந்தது)]. ஏடிஆர் (ATR) சர்வதேச மாநாடுகளில் தொலைபேசி பதிவுகள் செய்வதற்கும் ஹோட்டல் விடுதி தொலைபேசியில் முன்பதிவு செய்வதற்கும் ஒரு ஒழுங்குமுறையை உருவாக்கி வந்தது. சற்று பின்னர் கார்னகி-மெலன் பல்கலைக்கழகம் (Carnegie-Mellon University) லெக்ஸ் வைபெலின் கீழ் ஜானஸ் திட்டம் (JANUS project) வந்தது; பின்னர் கார்ல்ஸ்ரூ பல்கலைக்கழகத்துடன் (University of Karlsruhe)

(ஜெர்மனி (Germany)) ஒத்துழைத்தது மற்றும் ஏடிஆர் (ATR) உடன் ஒரு கூட்டமைப்பு சி-ஸ்டார் (பேச்சு மொழிபெயர்ப்பு மேம்பட்ட ஆராய்ச்சிக்கான கூட்டமைப்பு) (Consortium for Speech Translation Advanced Research (C-STAR)). ஜானஸ் திட்டமும் பயணத் திட்டமிடலில் கவனம் செலுத்தியது; ஆனால் இந்த ஒழுங்குமுறை உடனடியாக விரிவாக்கக்கூடிய வகையில் வடிவமைக்கப்பட்டது. ஏடிஆர் மற்றும் சி-ஸ்டார் திட்டங்கள் இரண்டும் தொடர்ந்தன. மூன்றாவது குறுகிய காலக் குழு அதன் முக்கிய மொழி திட்டத்தின் (Core Language project) (அல்ஷாவி (Alshawi) 1992) ஒரு பகுதியாக எஸ்ஆர்ஐ-ஆல் (SRI) (கேம்பிரிட்ஜ், யுகே) அமைக்கப்பட்டது; மற்றும் 'அரை-தருக்க வடிவங்கள்' ('quasi-logical forms') வழியாக ஸ்வீடிஷ்-ஆங்கில மொழிபெயர்ப்பு ஆயப்பட்டது. மிகப் பெரிய அளவில் நான்காவது பேச்சு மொழித் திட்டமான வெர்ப்மொபில் (Verbmobil) வொல்ப்காங் வால்ஸ்டரால் (Wolfgang Wahlster) இயக்கப்பட்டது மற்றும் 1993 முதல் 2000ஆம் ஆண்டு வரை பல பல்கலைக்கழகங்களில் ஜெர்மன் அரசாங்கத்தால் (German government) நிதியளிக்கப்பட்டது.. வெர்ப்மொபிலின் நோக்கம், ஆங்கிலம் சரளமாகத் தெரியாத ஜெர்மனியர்களும் ஜப்பானியர்களும் நேருக்கு நேர் ஆங்கில மொழி வணிகப் பேச்சுவார்த்தைகளைச் செய்ய வேண்டி ஒரு கடத்த இயலும் உதவிக் கருவியை (transportable aid) உருவாக்குவதாகும். யூரோட்ராவைப் போல (மேலே உள்ள பிரிவு 2.7), அடிப்படை குறிக்கோள் சாதிக்கப்படவில்லை என்றாலும் உரையாடல் மற்றும் பேச்சு மொழிபெயர்ப்புக்கான திறமையான வழிமுறைகளின் உருவாக்கமும் இந்தத் துறையில் ஜெர்மனியில் உயர்தர ஆராய்ச்சி குழுக்களை நிறுவுவதும் குறிப்பிடத்தக்க வெற்றிகளாகக் கருதப்பட்டன.

மிக அண்மைக்காலத்தில் ஆங்கிலம், பிரஞ்சு, ஜப்பானிய ((MedSLT) மொழிகளுக்கிடையில் மருத்துவர்-நோயாளி தகவல்தொடர்புக்காக பேச்சு மொழி திட்டங்கள் நிறுவப்பட்டன. இது மேற்குறிப்பிட்ட கேம்பிரிட்ஜ் மற்றும் ஸ்வீடனில் எஸ்ஆர்ஐ ஆய்வு அடிப்படையில் அமையும். இது ஆங்கிலம், பிரஞ்சு, ஜெர்மன், இத்தாலியன் (NESPOLE!) மொழிகளுக்கு இடையில் வணிகக் கருத்துப்பரிமாற்றம் மற்றும் சுற்றுலாவுக்கானதாகும். இது கற்றலான் (Catalan) மற்றும் ஸ்பானிஷ் (Spanish) (அதன் ஃபேம் துணைத் திட்டத்தில் (FAME ancillary project)) மொழிகளையும் உட்படுத்தும். எளியது யுஎஸ் இராணுவத்திற்காக (ஃப்ரேஸ்லேட்டர் (Phraselator) உருவாக்கப்பட்ட 'குரல் மொழிபெயர்ப்பி' ('voice translator')

ஆகும்; இது இந்தி (Hindi), தாய் (Thai), இன்ந்தோனீசியன் (Indonesian), பாஷ்டோ (Pashto), அரபிக் (Arabic) போன்ற மொழிகளில் பேச்சு வெளியீட்டைக் கொண்ட ஒருவகைத் தொடர்நூல் (phrasebook) ஆகும்.

பேச்சுமொழி மொழிபெயர்ப்பில் ஆர்வத்துடன் தளர்வாக தொடர்புடைய பல ஒழுங்குமுறைகள் தொலைக்காட்சி தலைப்புகளின் (அல்லது துணைத்தலைப்புகள்) மொழிபெயர்ப்பிற்காக உருவாக்கப்பட்டுள்ளன - ஆங்கிலத்திலிருந்து ஸ்பானிஷ் மற்றும் போர்த்துகீசிய மொழிகளில், ஆங்கிலத்திலிருந்து கொரிய மொழியில், ஆங்கிலத்திலிருந்து பிரஞ்சு மற்றும் கிரேக்க மொழிகளில் போன்றன. பணியின் குறிப்பிட்ட கட்டுப்பாடுகள் பேசும் மொழியைக் கையாள்வதை (பெயர்த்து எழுத்தப்பட்டு குறைக்கப்பட்டது) உட்படுத்துவது மட்டுல்ல ஆனால் திரைகளில் இட வரம்புகள், எழுதப்பட்ட பனுவல்களைப் புரிந்துகொள்வது பேச்சைக் காட்டிலும் மெதுவாக இருக்கும் என்ற உண்மையும் ஆகும்.

2.9.4. கலப்பின ஒழுங்குமுறைகள் (Hybrid systems)

கடந்த பதின்ம ஆண்டில் வழிமுறைகளின் விரிவாக்கம் மற்றும் தானியங்கி மொழிபெயர்ப்பு செயல்முறைகளுக்குப் புதிய பயன்பாடுகளை அறிமுகப்படுத்துதல் என்பன ஒரு ஒற்றை அணுகுமுறையைப் பின்பற்றுவதற்கான வரம்புகளை எடுத்துக்காட்டுகின்றன. கடந்த காலத்தில் மொழிபெயர்ப்பின் சிக்கல்கள், முடிவுறாத விளைவுகள் அல்லது எல்லைக்குட்பட்ட பயன்பாடுகள் கொண்ட இயந்திரமொழிபெயர்ப்பை ஒரு குறிப்பிட்ட கோட்பாடு அல்லது குறிப்பிட்ட முறைக்கு சோதனைத்தளம் என்று கண்ட ஆராய்ச்சியாளர்களால் தொடங்கப்பட்டன. நல்ல தரமுள்ள தானியங்கி மொழிபெயர்ப்பை அடைவதற்கு எந்த ஒரு முறையும் இருக்க முடியாது என்பது இப்போது பரவலாக அங்கீகரிக்கப்பட்டுள்ளது; மற்றும் எதிர்கால மாதிரிகள் ‘கலப்பினங்களாக’ (‘hybrids’) இருக்கும்; இது விதிமுறை அடிப்படையிலான, புள்ளியியல் அடிப்படையிலான மற்றும் எடுத்துக்காட்டு அடிப்படையிலான நெறிமுறைகளில் சிறந்தவற்றை இணைக்கிறது,

ஒரு அணுகுமுறை, இணையான இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளை இயக்குவது மற்றும் வெளியீடுகளை இணைப்பது என்ற ஆலோசனை ஆகும்: ‘பன்முக இயந்திரம்’ என

அழைக்கப்படுகிற (‘multi-engine’) ஒழுங்குமுறை - கார்னகி-மெலன் பல்கலைக்கழகத்தின் (Carnegie-Mellon University) ஒரு குழு, அறிவு அடிப்படையிலான மற்றும் எடுத்துக்காட்டு அடிப்படையிலான ஒழுங்குமுறைகளின் சேர்க்கைகள் குறித்து ஆராய்ந்துள்ளது. மிகவும் பொதுவாக, கலப்பினங்கள் தற்போது எஸ்எம்டி (SMT) அல்லது இபிஎம்டி-இன் (EBMT) புள்ளியியல்சார் முறைகளைச் சில மொழியியல் அடிப்படையிலான முறைகளுடன் (linguistics-based methods) (விதி அடிப்படையிலான அணுகுமுறைகளில் இருந்து) இணைக்கும் அமைப்புகள் என கருதப்படுகின்றன, குறிப்பாக உருபனியல் மற்றும் தொடரியல் பகுப்பாய்விற்கு. ஒரு எடுத்துக்காட்டு மைக்ரோசாப்டில் (டோலன் மற்றும் பலர் (Dolan et al. 2002)) ஆராய்ச்சி. இருப்பினும் தரவுத்தொகுதி அடிப்படையிலான மற்றும் விதி அடிப்படையிலானவற்றை இணைக்க வேறு வழிகள் உள்ளன; தேசிய சிங்-ஹுவா பல்கலைக்கழகத்தில் (National Tsing-Hua University) (தைவான் (Taiwan)) (சாங் & சு (Chang & Su 1997)) ஆராய்ச்சி எடுத்துக்காட்டு ஆகும்: சொல்சார் மற்றும் தொடரியல் விதிகள் புள்ளியியல் ரீதியாக தரவுத்தொகுதியிலிருந்து பெறப்படலாம் மற்றும் வெளியீட்டிலிருந்து பின்னூட்டத்தால் மேம்படுத்தப்படும்.

2.9.5. மதிப்பீடு (Evaluation)

இயந்திரமொழிபெயர்ப்பு மதிப்பீடு (MT evaluation) ஆராய்ச்சி நடவடிக்கைகளின் முக்கிய மற்றும் தீவிரமான பகுதியாக மாறியுள்ளது. 1990களில் பல பட்டறைகள் இயந்திரமொழிபெயர்ப்பை மதிப்பிடுவதில் உள்ள சிக்கல்களுக்குக் குறிப்பாக அர்ப்பணிக்கப்பட்டன, எ.கா. பால்கெடல் (Falkedal) 1994, வாஸ்கான்செலோஸ் (Vasconcellos) 1994 மற்றும் பல இயந்திரமொழிபெயர்ப்பு மாநாடுகளுடன் இணைக்கப்பட்ட பட்டறைகள். ஜப்பான் மின்னணு தொழில் மேம்பாட்டு சங்கத்தால் (Japan Electronic Industry Development Association (JEIDA) 1992)) உருவாக்கப்பட்ட நெறிமுறைகள் து மற்றும் ARPA (பின்னர் DARPA) ஆதரவு திட்டங்களின் மதிப்பீடுக்காக வடிவமைக்கப்பட்டவை குறிப்பாகச் செல்வாக்கு செலுத்தியது (ARPA 1994); மற்றும் இயந்திரமொழிபெயர்ப்பு மதிப்பீடு கணினி மொழியியலின் பிற பகுதிகளின் மதிப்பீட்டிற்கும் இயற்கையான மொழி ஆய்வின் மொழியியல் மற்றும் பிற பயன்பாடுகளுக்கும் குறிப்பிடத்தக்க தாக்கங்களை ஏற்படுத்தியது. ஆரம்பத்தில், இயந்திரமொழிபெயர்ப்பு தரத்தின் பெரும்பாலான நடவடிக்கைகள் புரிந்துகொள்ளுதல்

(comprehensibility), புத்திசாலித்தனம் (intelligibility), சரளம் (fluency), துல்லியம் (accuracy) மற்றும் பொருத்தம் (appropriateness) போன்ற காரணிகளின் மனித மதிப்பீடுகளால் நிகழ்த்தப்பட்டது. ஆனால் அத்தகைய மதிப்பீட்டு முறைகள் நேரம் மற்றும் முயற்சியி அடிப்படையில் விலை உயர்ந்தவை; எனவே தானியங்கி (அல்லது அரை தானியங்கி) முறைகளை உருவாக்க குறிப்பாக 2000 முதல் முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன.

புள்ளியியல் அடிப்படையிலான இயந்திரமொழிபெயர்ப்பு மாதிரிகள் (statistics-based MT models) (எஸ்எம்டி, மேலே பிரிவு 2.9.1) வளர்ச்சியின் ஒரு முக்கியமான விளைவு உண்மையில் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் தானியங்கி மதிப்பீட்டிற்கு புள்ளிளியியல்சார் ஆய்வின் பயன்பாடாகும். முதல் அளவீடு ஐபிஎம் குழுவிலிருந்து பி.எல்.இ.யு, (புளூ) (BLEU) பின்னர் என்.ஐ.எஸ்.டி. (நிஸ்ட் (NIST)) (தரநிலைகள் மற்றும் நுட்பங்களுக்கு தேசிய நிறுவனம்) (National Institute for Standards and Techniques (NIST)). இந்த இரண்டு நடவடிக்கைகளும் மனிதன் உற்பத்திசெய்த மொழிபெயர்ப்புகள் [‘குறிப்புப் பனுவல்கள்’ (‘reference texts’)] என அழைக்கப்படுகின்றன] கிடைப்பதை அடிப்படையாகக் கொண்டவை. இயந்திரமொழிபெயர்ப்பு அமைப்பிலிருந்து வெளியீடு கூடுதல் ‘குறிப்புப் பனுவல்களுள்’ ஒன்றுடன் ஒப்பிடப்படுகிறது; இயந்திரமொழிபெயர்ப்பு பனுவல்கள் ஒரே மாதிரியானவை அல்லது சொல் வரிசைகளின் அடிப்படையில் ‘குறிப்பு’க்கு மிக நெருக்கமானவை அதிக மதிப்பெண் பெறும்; தனிப்பட்ட சொல் நிகழ்வுகளில் அல்லது சொல் வரிசைகளில் பெரிதும் வேறுபடுகின்றன இயந்திரமொழிபெயர்ப்பு பனுவல்கள் குறைவாக மதிப்பெண் பெறும். அளவீடுகள் விதிமுறை அடிப்படையிலான ஒழுங்குமுறைகளை எஸ்எம்டி (SMT) ஒழுங்குமுறைகளை விடக் குறைவாக மதிப்பிடுகின்றன; இருப்பினும் முந்தையவை பெரும்பாலும் மனித வாசகர்களுக்கு மிகவும் ஏற்றுக்கொள்ளத்தக்கவையாக இருக்கின்றன. இதன் விளைவாக ஒரு குறிப்பிட்ட ஒழுங்குமுறை (SMT அல்லது EBMT) காலப்போக்கில் மேம்பட்டதா அல்லது இல்லையா என்பதைக் கண்காணித்தலுக்கு அவற்றின் மதிப்பை மறுப்பதற்கில்லை; ஆனால் ஒப்பீட்டு இயந்திரமொழிபெயர்ப்பு மதிப்பீடுகளுக்கான இந்த அளவீடுகளின் பொதுவான மதிப்பு குறித்து அதிக சந்தேகம் உள்ளது; கூடுதல் பொருத்தமான மற்றும் உணர்திறன் அளவீடுகளைத் தேடுவது தீவிரமடைந்துள்ளது.

2.10. 1990 முதல் செயல்பாட்டு மற்றும் வணிக ஒழுங்குமுறைகள்

1990 முதல் செயல்பாட்டு மற்றும் வணிக ஒழுங்குமுறைகள் (Operational and commercial systems since 1990) என்ற தலைப்பின் கீழ் ஹட்சின்ஸ் (Hutchins 2015) கூறியுள்ள செய்திகள் இங்கு தரப்பட்டுள்ளன. இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் பயன்பாடு 1990களில் துரிதப்படுத்தப்பட்டது. வணிக முகவர் நிலையங்கள் (commercial agencies), அரசு சேவைகள் (government services) மற்றும் பன்னாட்டு நிறுவனங்கள் (multinational companies) இவற்றில் பயன்பாட்டின் அதிகரிப்பு மிகவும் குறிக்கப்பட்டுள்ளது. இங்கு மொழிபெயர்ப்புகள் பெரிய அளவில் தயாரிக்கப்படுகின்றன; முதன்மையாக தொழில்நுட்ப ஆவணங்கள். முக்கிய சட்டக ஒழுங்குமுறைகளுக்கு இது முக்கிய சந்தையாக இருந்தது: சிஸ்ட்ரான் (Systran), லோகோஸ் (Logos), மெட்டல் (METAL) மற்றும் அட்லஸ் (ATLAS). எல்லாவற்றிலும் மொழிபெயர்ப்புகள் பெரிய தொகுதிகளில் தயாரிக்கப்படும் நிறுவல்கள் உள்ளன. ஏற்கனவே 1995இல் ஒரு ஆண்டிற்கு 300 மில்லியனுக்கும் அதிகமான சொற்கள் அத்தகைய நிறுவனங்களால் மொழிபெயர்க்கப்பட்டுள்ளதாக மதிப்பிடப்பட்டது.

மென்பொருள் ஓரிடமாக்கல் (software localisation) துறை வேகமாக வளர்ந்து வரும் ஒரு பகுதிகளில் ஒன்றாகும். இங்கே புதிய மென்பொருள் தொடங்கும் நேரத்தில் பல மொழிகளில் ஆதரிக்கும் ஆவணங்கள் (supporting documentation) கிடைக்கவேண்டும் என்பதே கோரிக்கை. மொழிபெயர்ப்பு விரைவாக செய்யப்பட வேண்டும் (மென்பொருள் விற்பனை செய்யப்படும் போது); ஆனால் ஆவணங்கள் ஒரு பதிப்பிலிருந்து மற்றொரு பதிப்பிற்கு மீள்நிகழும் தகவல்களைக் கொண்டுள்ளன. வெளிப்படையான தீர்வு இயந்திரமொழிபெயர்ப்பு மற்றும் மிக சமீபத்தில் மொழிபெயர்ப்பு பணிநிலையங்களில் (translation workstations) மொழிபெயர்ப்பு நினைவுகள். சமீபத்திய தொடர்புடைய வளர்ச்சி, நிறுவனத்தின் தளங்களில் வலைப்பக்கங்களை ஓரிடமாக்குவதாகும் - மீண்டும், தேவையானது உடனடி முடிவுகள் மற்றும் தகவல்களின் மீள்நிகழ்வு உள்ளது. வலைத்தள ஓரிடமாக்கத்திற்கான மொழிபெயர்ப்பு மென்பொருள் தயாரிப்புகளை விட மிக வேகமாக வளர்ந்து வருகிறது, இப்போது வலைத்தள உருவாக்குநர்களை (website developers) ஆதரிக்க கணினி கருவிகள் பல உள்ளன, எ.கா. ஐபிஎமின் வெப்ஸ்பியர் (IBM's WebSphere).

1990களில் குறிப்பிட்ட பொருண்மைக் களங்கள் மற்றும் பயனர்களுக்கான ஒழுங்குமுறைகளின் வளர்ச்சியும் விரைவாக விரிவடைந்தது - பெரும்பாலும் கட்டுப்படுத்தப்பட்ட

மொழிகளுடன் (controlled languages) மற்றும் குறிப்பிட்ட துணை மொழிகளின் அடிப்படையில். கட்டுப்படுத்தப்பட்ட மொழிகள் மொழிபெயர்ப்பின் போது பொருண்மைமயக்கநீக்கச் சிக்கல்களைக் குறைக்கச் சொற்றொகை (அங்கீகரிக்கப்பட்ட சொற்களின் தேர்வு அல்லது தனிப்பட்ட புலன்களின் பயன்பாடு) மற்றும் இலக்கணத்தின் (குறிப்பிட்ட ஆவண வகைகள் அல்லது பொருண்மைக் களங்களுக்குப் பொருத்தமான நடை விதிமுறைகள்) கட்டுப்பாடுகளை உட்படுத்தும். வோல்மேக் லிங்வேர் சேவைகளால் (Volmac Lingware Services) துணி நிறுவனம் (textile company), காப்பீட்டு நிறுவனம் (insurance company) இவற்றிற்கு வேண்டியும் மற்றும் விமான பராமரிப்பு கையேடுகளை (aircraft maintenance manuals) மொழிபெயர்க்கவும் ஒழுங்குமுறைகள் உருவாக்கப்பட்டன. காப் ஜெமினி இன்னொவஷன் (Cap Gemini Innovation) இராணுவ தொலை அச்சுச் செய்திகளை (military telex messages) மொழிபெயர்க்க ஒரு ஒழுங்குமுறையை உருவாக்கியது; ஜப்பானில், சி.எஸ்.கே. (CSK) நிதி மற்றும் பொருளாதாரத் துறையில் அதன் சொந்த ஒழுங்குமுறையை உருவாக்கியது; மற்றும் ஜப்பானிய செய்தி ஒளிபரப்பப்புகளை ஆங்கிலத்தில் மொழிபெயர்ப்பதற்கான ஒரு ஒழுங்குமுறை என்.எச்.கே. (NHK) உருவாக்கியது. லாண்ட் நிறுவனம் (LANT company) (பின்னர் எக்ஸ்ப்ளேனேஷன் பி.வி. (Xplanation b.v.)) பழைய மெட்டல் ஒழுங்குமுறையை (METAL system) அடிப்படையாகக் கொண்ட கட்டுப்படுத்தப்பட்ட மொழி இயந்திரமொழிபெயர்ப்பு (controlled language MT system) ஒழுங்குமுறைகளை உருவாக்கியது. எஸ்டீம் நிறுவனம் (ESTeam company) [கிரேக்கத்தை (Greece) தளமாகக் கொண்டது] ஆரம்பத்தில் உற்பத்தியாளர்களின் உபகரணக் கூறுகளின் (equipment components) பட்டியல்களின் 'கட்டுப்படுத்தப்பட்ட மொழியில்' கவனக்குவிப்பு செய்தது; ஆனால் பின்னர் பனுவல் போன்ற ஆவணங்களுக்கு விரிவாக்கப்பட்டது; பல நிறுவனங்களுக்கான வாடிக்கையாளர் ஒழுங்குமுறைகள் (customer systems) உருவாக்கப்பட்டன (மற்றும் சமீபத்தில் ஐரோப்பிய ஒன்றியம் உட்பட). கட்டுப்படுத்தப்பட்ட அனைத்து மொழி ஒழுங்குமுறைகளிலும் மிகவும் வெற்றிகரமாக தொடர்வது முன்னர் குறிப்பிடப்பட்ட ஸ்மார்ட் கார்ப்பரேஷனின் (Smart Corporation) 'ஸ்மார்ட் மொழிபெயர்ப்பாளர்' ('Smart Translator') ஆகும்.

1990களின் தொடக்கத்திலிருந்து தனிப்பட்ட கணினிகளுக்கான பல ஒழுங்குமுறைகள் தோன்றின. தனிநபர் கணினிகளின் அதிகரிக்கும் கணக்கீட்டு சக்தி மற்றும் சேமிப்பக திறன்கள் இந்த வணிக ஒழுங்குமுறைகளை 1980கள் மற்றும் அதற்கு முந்தைய பெரும்பாலான மெயின்பிரேம் அமைப்புகளுக்கு சமமானவைகளாக உருவாக்கின; மற்றும் பல சந்தர்ப்பங்களில் மேலும் சக்திவாய்ந்தவையாக உருவாக்கின. இருப்பினும் மொழிபெயர்ப்பு தரத்தில் சமமான

முன்னேற்றம் ஏற்படவில்லை. கிட்டத்தட்ட அனைத்தும் பழைய இடமாற்ற அடிப்படையிலான (அல்லது 'நேரடி மொழிபெயர்ப்பு') மாதிரிகளின் அடிப்படையில் ஆனவை; சிலவற்றில் கணிசமான மற்றும் நன்கு நிறுவப்பட்ட அகராதிகள் உள்ளன; பெரும்பாலான விற்பனையாளர்கள் பல்வேறு அறிவியல் மற்றும் தொழில்நுட்ப பொருண்மைகளுக்கான சிறப்பு அகராதிகள் வழங்கினாலும் பெரும்பாலான முயற்சிகள் பொது நோக்க ஒழுங்குமுறைகளாக செயல்பட முயற்சித்தன. கிட்டத்தட்ட எல்லா நிகழ்வுகளிலும் அமைப்புகள் மூன்று அடிப்படை பதிப்புகளில் விற்கப்படுகின்றன: பொதுவாக வாடிக்கையாளர்-சேவையக கட்டமைப்புகளில் (client-server configurations) இயங்குகின்ற பெரிய நிறுவனங்களுக்கான ஒழுங்குமுறைகள் ('நிறுவன' ஒழுங்குமுறைகள் ('enterprise' systems)), சுந்தந்திரமான மொழிபெயர்ப்பாளர்களுக்கான ஒழுங்குமுறைகள் ('தொழில்முறை' ஒழுங்குமுறைகள் ('professional' systems)) மற்றும் மொழிபெயர்ப்பாளர் அல்லாதவர்களுக்கான ஒழுங்குமுறைகள் ('வீட்டுப் பயன்பாடு' ('home use')).

தனிப்பட்ட கணினிகளில் பரவலாக விற்கப்படும் ஆரம்பகால அமைப்புகளில் இரண்டு, பிசி-மொழிபெயர்ப்பாளரும் (PC-Translator) [மொழியியல் தயாரிப்புகள், டெக்சாஸ்-இலிருந்து (from Linguistic Products, Texas] பவர் மொழிபெயர்ப்பாளரும் (Power Translator) [குளோபலிங்கிலிருந்து (from Globalink)] ஆகும். குளோபலிங்க் முதலில் மைக்ரோடாக் (MicroTac) உடன் [மொழி உதவித் தொடரின் தயாரிப்பாளர் (producer of the Language Assistant series)] இணைக்கப்பட்டது, பின்னர் லெர்னவுட் மற்றும் ஹவுஸ்பி (Lernout & Hauspie) ஆகியோரால் வாங்கப்பட்டது. சிஸ்ட்ரானின் அசல் மெயின்பிரேம் ஒழுங்குமுறைகள் இப்போது தனிநபர் கணினிகளில் நிறுவனங்களின் பயன்பாட்டிற்கான பதிப்புகளில் மட்டுமல்ல, தொழில்முறை மொழிபெயர்ப்பாளர்களுக்கும் வீட்டு உபயோகத்துக்கும் சந்தைப்படுத்தப்படுகின்றன. பான் அமெரிக்கன் ஹெல்த் ஆர்கனைசேஷனிலிருந்து (Pan American Health Organization) இரண்டு மெயின்பிரேம் அமைப்புகள் (SPANAM and ENGSPAN) சுதந்திர மொழிபெயர்ப்பாளர்களுக்காக பிசி மென்பொருளாக இப்போது கிடைக்கிறது. முதன்மையாக நிறுவனங்கள் மற்றும் தொழில்முறை பயனர்களுக்கு மெட்டல் அமைப்பு ஜி.எம்.எஸ்-ஆல் (GMS) [இப்போது செயில் லேப்ஸ் (Sail Labs)] தழுவப்பட்ட T1 ஒழுங்குமுறையாக லாங்கன்ஷெய்டால் (Langenscheidt) விற்கப்பட்டது (முக்கியமாக ஒரு 'வீடு' ஒழுங்குமுறையாக) மற்றும் காம்ப்ரெண்டியம் ஒழுங்குமுறையாக (Comprendium system) சேல் லேப்ஸால் (Sail Labs) விற்கப்பட்டது. ஐபிஎம்மின்

(IBM) எல்எம்டி ஒழுங்குமுறை (LMT system), தனிப்பட்ட மொழிபெயர்ப்பாளர் பி.டி ஆகக் (Personal Translator PT) [ஆரம்பத்தில் ஐ.பி.எம் மற்றும் வான் ரைன்பாபென் & புஷ் (von Rheinbaben & Busch) ஆகியோரால் விற்கப்பட்டது, இப்போது லிங்குடெக் ஜி.எம்.பி.எச். (Linguatec GmbH)-ஆல் விற்கப்பட்டது] குறைக்கப்பட்டது. பழைய வீட்னர் (Weidner system) ஒழுங்குமுறையிலிருந்து பெறப்பட்ட டிரான்ஸ்ஸெண்ட் சிஸ்டம் (Transcend system) [வெளிப்படையான மொழி (Transparent Language)] [இன்டர்கிராஃப் வழியாக (via Intergraph)] இப்போது எளிய மொழிபெயர்ப்பாளர் (Easy Translator) (ஒரு 'வீட்டு' அமைப்பு) மற்றும் நிறுவன மொழிபெயர்ப்பு சேவையகம் (Enterprise Translation Server) [இரண்டும் முதலில் வெளிப்படையான மொழி இன்க். இலிருந்து (from Transparent language Inc.)] இவற்றுடன் இணைந்து எஸ்.டி.எல். நிறுவனத்தால் (SDL company) [மொழிபெயர்ப்பாளரின் பணிநிலையத்தின் உருவாக்குநர்கள் (developers of a translator's workstation)] விற்கப்படுகிறது.

முன்னாள் சோவியத் யூனியனில் இருந்து ஸ்டைலஸ் (பின்னர் புரோஎம்டி என மறுபெயரிடப்பட்டது) மற்றும் PARS என்ற ரஷ்ய மற்றும் ஆங்கில மொழிபெயர்ப்புகான சந்தைப்படுத்தும் ஒழுங்குமுறைகளும் மற்றும் பிற மொழிகளுக்கான ஒழுங்குமுறைகளும் (பிரெஞ்சு, ஜெர்மன், உக்ரேனிய, முதலியன) வந்தன. மேற்கு ஐரோப்பாவில் புரோஎம்டி அமைப்புகள் ஒரு காலத்திற்கு ரெவர்சோ தொடராக சந்தைப்படுத்தப்பட்டன சாஃப்டிஸிமோ நிறுவனம். ஐரோப்பாவிலிருந்து பிற பிசி அடிப்படையிலான அமைப்புகள் பின்வருமாறு: இத்தாலிய மொழியில் மொழிபெயர்க்க பெட்ரா மற்றும் ஆங்கிலம்; டேனிஷ்-ஆங்கிலம், பிரஞ்சு-ஆங்கிலம் மற்றும் ஆங்கிலம்-ஸ்பானிஷ் மொழிகளுக்கான விங்கர் அமைப்பு (இப்போது இல்லை கிடைக்கும்); மற்றும் டிரான்ஸ்மார்ட், கீலிகோன் அமைப்பின் வணிக பதிப்பு (முதலில் உருவாக்கப்பட்டது நோக்கியா), பின்னிஷ்-ஆங்கில மொழிபெயர்ப்புக்கு.

ஜப்பானிய மற்றும் ஆங்கிலத்திற்கான பல ஒழுங்குமுறைகள் ஜப்பானிய நிறுவனங்களிலிருந்து தொடர்ந்து தோன்றின; உண்மையில், ஒழுங்குமுறைகளின் திகைப்பூட்டும் வரலாறு உள்ளது; வெற்றிகரமாக விற்பனை குறுகிய காலத்திற்கு பின்னர் மறைந்துவிடும். நீண்ட காலம் நீடித்தவைகளில் புஜித்சுவின் தயாரிப்புகள் (அட்லாஸ் (ATLAS)), ஹிட்டாச்சி (HICATS), தோஷிபா (Toshiba) (ஆரம்பத்தில் ASTRANSAC என அழைக்கப்பட்டது), NEC [கிராஸ்ரோட் (Crossroad), முன்பு பிவோட் (Pivot)], குறுக்கு மொழி (Cross Language) (முன்னர் நோவா: டிரான்சர்

தொடர் (formerly Nova: the Transer series)), மற்றும் ஷார்ப் (Sharp) (பவர் எ/ஜே, இப்போது ஹொனியாகு கோரே-இப்போன் (Power E/J, now Honyaku Kore-Ippon)) என்பன அடங்கும். ஆனால் நல்ல தரமான அமெரிக்க தயாரிப்புகளும் உள்ளன: மொழி பொறியியல் கார்ப்பரேஷனில் இருந்து (Language Engineering Corporation) லோகோவிஸ்டா ஒழுங்குமுறை (LogoVista system) [பின்னர் லோகோமீடியாவால் (LogoMedia) எடுத்துக்கொள்ளப்பட்டது], நியோகார் டெக்னாலஜிஸிலிருந்து (Neocor Technologies) [லெர்னவுட் & ஹவுஸ்பியால் (Lernout & Hauspie) வாங்கப்பட்டது, இப்போது கிடைக்காது] சுனாமி (Tsunami) மற்றும் டைபூன் (Typhoon) ஒழுங்குமுறைகள், அத்துடன் வெப்ஸ்பியர் மற்றும் சிஸ்ட்ரானிலிருந்து ஒழுங்குமுறைகள். ஆங்கிலத்திலிருந்து கொரிய மொழிக்கு ஒழுங்குமுறைகள் 1990களின் பிற்பகுதியில் மட்டுமே வந்தன (எ.கா. சிஸ்ட்ரான், லோகோமீடியா, டிரான்ஸ்பியர் (Systran, LogoMedia, TranSphere) போன்றவை).

அரபு மொழியில் இப்போது பல ஒழுங்குமுறைகள் உள்ளன (குறிப்பாக சக்ர், சிமோஸ் மற்றும் ஆப் டெக் ஒழுங்குமுறைகள் (Sakhr, Cimos and AppTek systems)); மற்றும் சீனமொழிக்கான அதிகரித்து வரும் எண்ணிக்கையிலான ஒழுங்குமுறைகள் (எ.கா. டிரான்ஸ்டார், லோகோமீடியா, சிஸ்ட்ரான், டிரான்ஸ்பியர் (Transtar, LogoMedia, Systran, TranSphere)) உள்ளன; அமெரிக்க அரசாங்க நிறுவனங்களின் கவனத்தை ஈர்ப்பது மிக சமீபத்திய ஆராய்ச்சியின் அடிப்படையில் இந்த இரண்டு மொழிகளுக்கான ஒழுங்குமுறைகளை உருவாக்குபவர்களுக்கு அதிக ஊக்கம் அளிக்கிறது.

ஆங்கிலம் தவிர மூலமாகவோ அல்லது இலக்காகவோ உள்ள பிற மொழி இணைகளுக்கான ஒழுங்குமுறைகள் குறைவாகவே உள்ளன. இருப்பினும், மேலே குறிப்பிட்டுள்ள பெரும்பாலான அமெரிக்க மற்றும் ஐரோப்பிய நிறுவனங்கள் பிரஞ்சு-ஜெர்மன், இத்தாலியன்-ஸ்பானிஷ், போர்த்துகீசிய-ஸ்பானிஷ் என்ற இணைகளுக்கு அத்தகைய ஒழுங்குமுறைகளை வழங்குகின்றன; மற்றும் ஜப்பானிய-சீன, ஜப்பானிய-கொரிய முதலியன இணகளுக்கு மொழிபெயர்ப்பு நல்கும் ஒழுங்குமுறைகளும் உள்ளன.

இந்த வணிக நடவடிக்கை இருந்தபோதிலும், இன்னும் பல மொழிகள் மோசமாகவே உள்ளன. ஆப்பிரிக்கா, இந்தியா மற்றும் தென்கிழக்கு ஆசியா இவற்றிலுள்ள பெரும்பாலான

மொழிகளுக்கு வணிக ஒழுங்குமுறைகளின் பற்றாக்குறை இன்னும் உள்ளது; இருப்பதையும் எளிதில் அணுக இயலாது.

மிகச் சமீபத்திய வளர்ச்சியானது மொழிபெயர்ப்பிற்கான வாக்கியங்களின் பேச்சு உள்ளீடு (அல்லது உரைகள்) மற்றும் விளைவுகளின் பேச்சு வெளியீடு வழிமுறையை பல பிசி ஒழுங்குமுறைகளில் இணைப்பதாகும். (இவை நிச்சயமாக இல்லை, உண்மையான பேச்சு மொழிபெயர்ப்பு அமைப்புகள், மேலே 9.3 இல் சுட்டிக்காட்டப்பட்டுள்ளபடி, இன்னும் செயல்பாட்டுக்கு வரவில்லை.) சந்தேகத்திற்கு இடமின்றி பேச்சு அறிதல் ஒழுங்குமுறைகளின் மேம்பாட்டு நம்பகத்தன்மை மற்றும் பயன்பாட்டின் எளிமை ஆகியவற்றை அடிப்படையாகக் கொண்ட இந்தத் தயாரிப்புகள் 1990களின் கடைசி ஆண்டுகளில் தோன்றத் தொடங்கின. மிக சமீபத்திய எடுத்துக்காட்டுகளில் ஒன்று ஐபிஎம்மில் இருந்து குரல்வழி மொழிபெயர்ப்பாளர் (ViaVoice Translator); இது அதன் எல்எம்டி தானியங்கி மொழிபெயர்ப்பு ஒழுங்குமுறை மற்றும் அதன் வெற்றிகரமான குரல்வழி சொல்வதெழுதும் ஒழுங்குமுறை (ViaVoice dictation system) இவற்றின் இணைப்பாகும்.

நிறுவனத்தின் கையகப்படுத்துதல், பழைய ஒழுங்குமுறைகளின் அழிவு மற்றும் வழக்கமான புதிய ஒழுங்குமுறைகளின் தோற்றம் மற்றும் பழைய ஒழுங்குமுறைகளின் புதிய பதிப்புகள் இவற்றின் எடுத்துக்காட்டுகள் வணிக மாற்றம் இயந்திர மொழிபெயர்ப்பு வரலாற்றின் ஆராய்ச்சி முன்னேற்றங்களின் ஒரு பகுதி என்பதை விளக்குகிறது மற்றும் பிற வணிகத்தைப் போன்று தோல்வி இயந்திர மொழிபெயர்ப்பின் அம்சமாகவும் உள்ளது. முதல் நிகழ்வு வீட்னர்/ பிரேவிஸ் (Weidner/Bravice) சரிவு (மேலே உள்ள பிரிவு 6). பல சிறிய நிறுவனங்கள் கடந்த இரண்டு பதின்ம ஆண்டுகளில் வந்து போயின. ஆனால் சந்தையில் தங்கள் நிலையைத் தக்க வைத்துக் கொள்ள முடியாத வெளிப்படையாக வெற்றிகரமான நிறுவனங்களின் உதாரணங்களும் உள்ளன. லோகோஸ் கார்ப்பரேஷன் (Logos Corporation) 1980களில் சிஸ்ட்ரான் இன்க் (Systran Inc) நிறுவனத்தை அதன் 'நிறுவன' ஒழுங்குமுறைகளுடன் எதிர்த்தது; ஆனால் 1990களில் பிற்பகுதியில் வேறொரு நிறுவனத்திற்கு (குளோபல் வேர்ட்ஸ் ஏஜி (globalwords AG)) விற்கப்பட்டது; இப்போது முற்றிலும் போய்விட்டது. காம்ப்ரெண்டியம் ஒழுங்குமுறை (Comprendium system) 1990களில் (பகுதி மெட்டல் ஒழுங்குமுறையின் அடிப்படையில்) பெரிய நிறுவனங்களுக்கு நல்ல தரமான ஒழுங்குமுறைகளை வழங்கியது; ஆனால் அதுவும்

போய்விட்டது. நன்கு விளம்பரப்படுத்தப்பட்ட (மோசமான) உதாரணம் லெர்னவுட் மற்றும் ஹாஸ்பியின் (Lernout & Hauspie) கதை. இந்த நிறுவனம் பேச்சு புரிதல் தயாரிப்புகளில் அதன் நற்பெயரை உருவாக்கியது. 1998இன் ஆரம்பத்தில் அது இயந்திரமொழிபெயப்பு களத்தில் விரிவாக்க முடிவு செய்தது; ஒரு பரவலான மொழிகளுக்கான ஒழுங்குமுறைகளை ஒரு தயாரிப்பாக (iTranslator) ஒருங்கிணைக்கும் நோக்கத்துடன் பல நிறுவனங்களை வாங்கியது (குளோபலிங்க், நியோகோர், ஆப் டெக், சாய்ல் லேப்ஸ், அல்லஜிக் Globalink, Neocor, AppTek, Sail Labs, AILogic). அதே நேரத்தில் அது மொழிபெயர்ப்பு சேவைகளில் (மெண்டெஸ் எஸ்.ஏ (Mendez SA)) ஆர்வங்களைப் பெற்றது மற்றும் பெல்ஜியத்தில் உள்ள அரசாங்க மூலங்களிலிருந்து நிதி உதவியைப் பெற்றது. ஆனால் நிறுவனம் அளவிற்கு அதிகமாகத் தன்னை நீட்டித்தது; நிதி சிக்கல்களில் சிக்கியது (சில முறைகேடுகள் உட்பட) மற்றும் 2001இன் பிற்பகுதியில் கலைக்கப்பட்டது. அது வாங்கியிஅ சில நிறுவனங்கள் தங்களை மீண்டும் நிலைநிறுத்திக் கொள்ள முடிந்தது (எ.கா. ஆப் டெக் மற்றும் சுருக்கமாக சாய்ல் லேப்ஸ் (AppTek and, briefly, Sail Labs)) ஆனால் மற்ற சந்தர்ப்பங்களில் சந்தை திறமையான மற்றும் முன்னர் நன்கு பெறப்பட்ட ஒழுங்குமுறைகளை இழந்தது (எ.கா. குளோபலிங்க் மற்றும் நியோகோர் (Globalink and Neocor) போன்றவை)).

இயந்திரமொழிபெயப்பு சந்தையில் இந்த பலவீனம் 2000க்குப் பிறகு எஸ்எம்டி (SMT) நெறி முறைகளின் அடிப்படையில் வணிக ரீதியாக ஒழுங்குமுறைகள் தோன்றுவது வரை ஏன் வெற்றிகரமாக இல்லை என்பதை விளக்கக்கூடும். நன்கு நிறுவப்பட்ட விதி அடிப்படையிலான முறைகளுக்கு மாறாக, ஒரு பதின்மாஆண்டிற்கு மேலாக மட்டுமே செயலில் உள்ளது என்பதால் எஸ்.எம்.டி ஆராய்ச்சியாக புள்ளியியல்சார் அணுகுமுறைகள் மிகவும் ஆபத்தானவை அல்லது முதிர்ச்சி அடையாதது. ஆயினும்கூட, இப்போது சந்தையில் SMT ஒழுங்குமுறைகள் உள்ளன (மொழி வீவரிலிருந்து (from Language Weaver)), கணிசமாக அரசாங்க ஆதரவை ஈர்த்த மொழி இணைகளுக்கு (அரபு-ஆங்கிலம், சீன-ஆங்கிலம், இந்தி-ஆங்கிலம், சோமாலி-ஆங்கிலம்) மற்றும் குறிப்பாக விதி அடிப்படையிலான அணுகுமுறைகளுக்கு கடினமான சவால்களை உருப்படுத்தம் செய்த மொழி இணைகளுக்கு. சமீபத்தில், மொழி வீவர் கூகிள் உடன் இணைந்தது, ஏஸ்டி (SMT) அணுகுமுறைகளுக்கு அதன் பரந்த உரை வளங்கள் குறிப்பாக பொருத்தமானவை - மீண்டும்,

தற்போது சோதனைக்கு உட்பட்ட மொழி இணைகள் அரபு-ஆங்கிலம், சீன-ஆங்கிலம், ஜப்பானிய-ஆங்கிலம் போன்றவை.

2.11. இணையத்தில் இயந்திர மொழிபெயர்ப்பு

ஹட்சின்ஸ் (Hutchins 2015) 'இணையத்தில் இயந்திர மொழிபெயர்ப்பு' (MT on the Internet) என்ற தலைப்பில் கூறிய செய்திகள் இங்கு தொகுத்துத் தரப்பட்டுள்ளன. 1990களின் நடுப்பகுதியில் இருந்து இணையம் இயந்திர மொழிபெயர்ப்பு வளர்ச்சியில் சக்திவாய்ந்த செல்வாக்கை செலுத்தியது. முதலாவதாக, வலைப்பக்கங்கள் (Web pages) மற்றும் மின்னணு அஞ்சல் செய்திகளை (electronic mail messages) இணையாநிலையில் (ஆஃப்லைனில் (offline)) (அதாவது பெற்ற போது அல்லது அனுப்புவதற்கு முன்) மொழிபெயர்ப்பிற்காக இயந்திர மொழிபெயர்ப்பு மென்பொருள் தயாரிப்புகள் தோன்றின. இதற்கு ஜப்பானிய நிறுவனங்கள் வழிவகுத்தன; அவை விரைவாக வேறு இடங்களில் பின்தொடரப்பட்டன. இரண்டாவதாக 1990களின் நடுப்பகுதியில் தொடங்கி பல இயந்திர மொழிபெயர்ப்பு விற்பனையாளர்கள் தேவைக்கேற்ப மொழிபெயர்ப்பிற்கான இணைய அடிப்படையிலான இணைநிலை (ஆன்லைன் (online)) மொழிபெயர்ப்பு சேவைகளை (Internet-based online translation services) வழங்கி வந்தனர். 1980களில் பிரான்சில் சிஸ்ட்ரானால் மினிடெல் நெட்வொர்க்கில் (Minitel network) வழங்கப்பட்ட சேவை முன்னோடியான சேவை ஆகும்; ஆனால் இந்த யோசனை 1995ஆம் ஆண்டில் டிரான்ஸ்ஸெண்ட் ஒழுங்குமுறையை (Transcend system) அடிப்படையாகக் கொண்ட ஒரு சோதனை சேவையை (trial service based) கம்ப்யூசர்வ் (CompuServe) அறிமுகப்படுத்திய வரை பரவலாக எற்கப்படவில்லை. விரைவில், பிரஞ்சு, ஜெர்மன் மற்றும் ஸ்பானிஷ் மொழிகளை ஆங்கிலத்திலிருந்து மொழிபெயர்க்க (பின்னர் பல மொழி இணைகளுக்கும்) சிஸ்ட்ரானின் பதிப்புகளை வழங்கும் ஆல்டாவிஸ்டா தளத்தில் (AltaVista site) இப்போது நன்கு அறியப்பட்ட சேவை பாபெல்ஃபிஷ் (Babelfish) தோன்றியது. அது பல பிற இணைநிலை (ஆன்லைன்) சேவைகளால் (அவற்றில் பெரும்பாலானவை இலவசமாக) தொடரப்பட்டது; எ.கா. இணைப்புநிலையில் (ஆன்லைனில்) சாஃப்டிஸிமோ (Softissimo) அதன் ரெவர்சோ ஒழுங்குமுறைகளின் (Reverso systems) ஆன்லைன் பதிப்புகளுடன், லோகோ விஸ்டா (LogoVista) மற்றும் பார்ஸின் (PARS) ஆன்லைன் பதிப்புகளுடன் லோகோமீடியா (LogoMedia). சில

சேவைகள் கூடுதல் விலையில் மனித மொழிபெயர்ப்பாளர்களால் (மறுபரிசீலனை செய்பவர்கள்) பிந்தைய திருத்தியமைத்தலை (post-editing) வழங்குகின்றன; ஆனால் பெரும்பாலான சந்தர்ப்பங்களில் எந்த வகையிலும் தொடாமல் முடிவுகள் வழங்கப்படுகின்றன. அவற்றில் பலவற்றை இயந்திர மொழிபெயர்ப்பு 'இணைய முகப்புகள் ('எம்டி போர்ட்டல்கள்' ('MT portals')) மூலம் அணுகலாம்; அதாவது சுதந்திர சேவைகள் (independent services) ஒன்று அல்லது அதற்கு மேற்பட்ட ஒழுங்குமுறை விற்பனையாளர்களிடமிருந்து மொழிபெயர்ப்பு ஒழுங்குமுறையின் ஒரு பரப்பெல்லையை வழங்குகிறது.

தடப்பட்ட பல மூலப் பனுவல்களின் பேச்சுவழக்குத் தன்மையையின் தவிர்க்க இயலாமையால் இணைநிலை இயந்திர மொழிபெயர்ப்பு (ஆன்லைன் எம்டி) சேவைகளின் மொழிபெயர்ப்பின் தரம் பெரும்பாலும் மோசமாக இருந்தது. ஆனால் இந்த சேவைகள் தகவல் நோக்கங்களுக்காகப் பயனர்களின் சொந்த மொழிகளில் உடனடி செப்பமற்ற மொழிபெயர்ப்புகளுக்கான ஒரு குறிப்பிடத்தக்க (மற்றும் வெளிப்படையாக ஏற்றுக்கொள்ளத்தக்க) கோரிக்கையைச் சந்தேகத்திற்கு இடமின்றி நிரப்பிக்கொண்டிருந்தன; 1960களில் முந்தைய மெயின்பிரேம் அமைப்புகள் வழங்கிய பெரும்பாலும் பின்னரும் பின் வந்த ஆண்டுகளிலும் புறக்கணிக்கப்பட்ட செயல்பாடு. அவற்றின் பரவலான பயன்பாடு மற்றும் இயந்திர மொழிபெயர்ப்பின் பொது 'பிம்பத்தில்' அவை வெளிப்படையான தாக்கத்தை, பெரும்பாலும் எதிர்மறைத் தாக்கத்தை ஏற்படுத்தினாலும் இணைநிலை இயந்திர மொழிபெயர்ப்பு (ஆன்லைன் எம்டி) சேவைகள் பெரும்பாலான இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியாளர்களால் பெரும்பாலும் புறக்கணிக்கப்பட்டுள்ளன (காஸ்பரி (Gaspari 2004)).

இயந்திர மொழிபெயர்ப்புக்கு ஒரு குறிப்பிட்ட சவால் என்னவென்றால் பயனர்களுக்கு நன்கு தெரியாத மொழிகளில் மொழிபெயர்க்க இணைநிலை (ஆன்லைன்) ஒழுங்குமுறைகளைப் பயன்படுத்துவது ஆகும். இணையத்தில் பயன்படுத்தப்படும் பெரும்பாலான மொழிகள் பேச்சுவழக்கு, பொருத்தமற்றவை, 'ஒழுங்கற்றவை', சுருக்கெழுத்துக்கள் மற்றும் சுருக்கங்கள், குறிப்புகள், துணுக்குகள், நகைச்சுவைகள் போன்றவை நிறைந்தவை; மின்னணு அஞ்சல் மற்றும் அரட்டை அறைகள் மற்றும் மொபைல் தொலைபேசிகளின் மொழிக்கு இது குறிப்பாக உண்மை ஆகும். இந்த வகையான மொழிப் பயன்பாடு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள்

உருவாக்கப்பட்ட அறிவியல் மற்றும் தொழில்நுட்ப பனுவல்களின் மொழியிலிருந்து பெரிதும் வேறுபடுகின்றன. இருப்பினும், சமீபத்தில் ஒரு இங்கிலாந்து நிறுவனமான டிரான்ஸ்லூஷன் (Translution) ஆங்கிலம், பிரஞ்சு, ஜெர்மன், இத்தாலியன் மற்றும் ஸ்பானிஷ் இவற்றிற்கு இடையே மின்னஞ்சல்களின் (அத்துடன் வலைப்பக்கங்கள்) ஆன்லைன் மொழிபெயர்ப்பிற்கான ஒரு ஒழுங்குமுறையை வெளியிட்டுள்ளது. கோரிக்கை கணிசமாக இருக்க வேண்டும்;மேலும் எதிர்காலத்தில் கூடுதல் ஒழுங்குமுறைகள் வரும் என்பதில் சந்தேகமில்லை.

இணையம் ஓரளவு குறைவான மோசமான நிறுவனங்களை மின்னணு அகராதிகள் (அல்லது சொற்றொடர் புத்தகங்கள்) ஆன்லைன் பதிப்புகளை 'மொழிபெயர்ப்பு ஒழுங்குமுறைகளாக வழங்க ஊக்குவித்துள்ளது. அத்தகைய தயாரிப்புகளைப் பயன்படுத்தும் முழு வாக்கியங்களையும் (மற்றும் பனுவல்களையும்) மொழிபெயர்க்கப் பயன்படுத்தும் எவரும் திருப்தியற்ற முடிவுகளைப் பெறுவர்; பயனர்களுக்கு இலக்கு மொழிகள் தெரியாவிட்டால் எதையும் அவர்கள் புரிந்துகொள்ள முடியாத அளவிற்கு அவர்கள் முடிவுகளை அறிந்திருக்க மாட்டார்கள். இத்தகைய அமைப்புகள் சந்தேகத்திற்கு இடமின்றி இயந்திர மொழிபெயர்ப்பின் ஒட்டுமொத்தக் கருத்துக்குத் தீங்கு விளைவிக்கும்; சில விற்பனையாளர்களும் சேவை வழங்குநர்களும் தங்கள் முழுமையான தானியங்கி ஒழுங்குமுறைகள் (ஆன்லைனிலோ அல்லாமலோ) எப்போதும் எச்சரிக்கையுடன் பயன்படுத்தப்பட வேண்டிய செப்பமற்ற பதிப்புகளை மட்டுமே உருவாக்குகிறது என்று வலியுறுத்துகின்றனர்.

2.12. முடிவுரை

பரவலாக மாறுபட்ட மொழிபெயர்ப்பைச் சந்திக்க பல்வேறு வகையான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் தேவை என்பது இப்போது தெளிவாகியுள்ளது. இதுவரை அடையாளம் காணப்பட்டவைகள் பெரிய நிறுவனங்களுக்கான பொதுவாக ஒரு கட்டுப்படுத்தப்பட்ட களத்திற்குள் பாரம்பரிய இயந்திர மொழிபெயர்ப்பு ஒழுமுறைகளை அடங்கும். மொழிபெயர்ப்பு கருவிகள் மற்றும் பணிநிலையங்கள் (இயந்திர மொழிபெயர்ப்பு தொகுதிகள் விருப்பங்களாக) தொழில்முறை மொழிபெயர்ப்பாளர்களுக்காக வடிவமைக்கப்பட்டுள்ளன; அவ்வப்போது மொழிபெயர்ப்புகளுக்கான மலிவான பிசி அமைப்புகள்; கண்காணிப்பு அல்லது தகவல் சேகரிப்பு நோக்கங்களுக்காக தோராயமான

சாராம்சங்கள் பெற ஒழுங்குமுறைகளின் பயன்பாடு; மின்னணு அஞ்சல் மற்றும் வலைப்பக்கங்களை மொழிபெயர்ப்பதற்கு இயந்திர மொழிபெயர்ப்பு பயன்பாடு; நிலைபேறுபற்ற செய்திகளை அறியப்படாததா மொழிகளிலிருந்து மொழிபெயர்க்க ஒருமொழிபேசுபவர்களுக்கான ஒழுங்குமுறைகள்; கட்டுப்படுத்தப்பட்ட களங்களில் பேச்சு மொழிபெயர்ப்புக்கான ஒழுங்குமுறைகள். இந்த தேவைகளில் சில பூர்த்தி செய்யப்பட்டுள்ளன அல்லது செயல்திறமான ஆராய்ச்சிக்கு உட்பட்டது, ஆனால் இன்னும் பல சாத்தியங்கள் உள்ளன, குறிப்பாக இயந்திர மொழிபெயர்ப்பை மொழி தொழில்நுட்பத்தின் பிற பயன்பாடுகளுடன் இணைத்தல் (தகவல் மீட்டெடுப்பு, தகவல் பிரித்தெடுத்தல், சுருக்கம், முதலியன). பல வகைகயான இயந்திர மொழிபெயர்ப்பு அமைப்புகள் மிகவும் பரவலாக அறியப்பட்டு பயன்படுத்தப்படுவதால், மொழிபெயர்ப்பு தேவைகளின் சாத்தியமான வரம்பு மற்றும் சாத்தியமான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மிகவும் வெளிப்படையாகி, இதுவரை கற்பனை செய்யப்படாத திசைகளில் ஆராய்ச்சி மற்றும் மேம்பாட்டைத் தூண்டும்.

இயல் 3

சில முக்கியமான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள்

3.1. அறிமுகம்

கணிப்பொறி வழி மொழி பெயர்ப்பினைத் தானியங்கி மொழிப்பெயர்ப்பு (Automatic Translation) என்றும் இயந்திர மொழிபெயர்ப்பு (Machine Translation) என்றும் கணிப்பொறி உதவியுடன் நடைபெறும் மொழிபெயர்ப்பு என்றும் கூறலாம். கணிப்பொறியின் வேகம், நினைவகத்திறன், பதிவு செய்துள்ள செய்திகளைத் தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளப்பெறக்கூடிய வசதி (Retrievel System) போன்றவைக் கணிப்பொறி வழி மொழி பெயர்ப்புப் பணியை ஊக்குவிக்க உதவின.

அரசியல் காரணங்களுக்காக அமெரிக்காவில் நடந்த ஆய்வுகள் ரஷ்யமொழி-ஆங்கிலம் இயந்திர மொழிபெயர்ப்புக்காகவும் ரஷ்யாவில் நடந்த ஆய்வுகள் ஆங்கில-ரஷ்யமொழி இயந்திர மொழிபெயர்ப்புக்காகவும் நிகழ்த்தப்பட்டது. கீழே தரப்பட்டுள்ள இயந்திர மொழிபெயர்ப்பு குறித்த வரலாற்றுச் செய்திகள் ஸ்லோகம் (Slocum, 1985) வெளியிட்ட A survey of machine translation: its history, current status, and future prospects என்ற கட்டுரையையும் ஜான் ஹட்சின்ஸ் மற்றும் ஹெரால்ட் எல். ஸொமெர்ஸ் (John Hutchins and Harold L. Somers 1992) என்போரால் எழுதப்பட்ட An Introduction to Machine Translation என்ற நூலையும் பயன்படுத்தி எழுதப்பட்டுள்ளது. எனவே இதில் தர்ப்பட்ட விளக்கங்கள் அவர்களின் காலகட்டத்திற்குப் பொருந்தும். அவற்றின் இன்றைய நிலமை (2019 ஆண்டின் நிலமை) இதில் விளக்கப்படவில்லை.

3.2. காட்: ஜார்ஜ்டவுண் தானியங்கு மொழிபெயர்ப்பு (GAT: GEORGETOWN AUTOMATIC TRANSLATION)

ஸொல்கம் (Solcum1985) தமது "இயந்திர மொழிபெயர்ப்பின் ஒரு ஆய்வு: அதன் வரலாறு, தற்போதைய நிலை மற்றும் எதிர்கால வாய்ப்புகள்" ("A survey of machine translation: its history,

current status, and future prospects") என்ற கட்டுரையில் ஜார்ஜ்டவுணின் காட் (Georgetown's GAT) என்பது குறித்து எழுதிய தகவலின் படி இப்பகுதி அமைகின்றது.

இயந்திர மொழிபெயர்ப்பின் தொடக்ககால ஆய்வுகளின் ஒரு இடமாக ஜார்ஜ்டவுண் பல்கலைக்கழகம் அமைந்தது. 1952-இல் தொடங்கப்பட்ட இவ்வாய்வுக்கு அமெரிக்க அரசாங்கத்தால் ஆதரவு அளிக்கப்பட்டது. ஜார்ஜ்டவுனின் கேட் அமைப்பு 1964இல் ஓக் ரிட்ஜ் தேசிய ஆய்வகத்தில் (Oak Ridge National Laboratory) அணுசக்தி ஆணையத்திற்கும் (Atomic Energy Commission) மற்றும் இஸ்ப்ரா, இத்தாலியில் ஐரோப்பாவின் தொடர்புடைய ஆராய்ச்சி வசதி யூரடோமுக்கும் (EURATOM) வழங்கப்பட்டுச் செயல்பாட்டுக்கு வந்தது. இரண்டு அமைப்புகளும் ரஷ்ய இயற்பியல் பனுவல்களை "ஆங்கிலத்தில்" மொழிபெயர்க்கப் பல ஆண்டுகளாகப் பயன்படுத்தப்பட்டன. மனித மொழிபெயர்ப்புடன் ஒப்பிடுகையில் வெளியீட்டுத் தரம் மிகவும் மோசமாக இருந்தது; ஆனால் இவற்றின் நோக்கம் ஆவணங்களை விரைவாக ஸ்கேன் செய்து அவற்றின் உள்ளடக்கம் மற்றும் ஆர்வத்தை தீர்மானிப்பதை நோக்கமாக் கொண்டதால் GAT ஒழுங்குமுறை மாற்று முறைகளைவிட (மெதுவான மற்றும் அதிக விலையுள்ள மனித மொழிபெயர்ப்பு அல்லது மோசமான எந்த மொழிபெயர்ப்பும் இல்லாமை) உயர்ந்ததாக இருந்தது. யுரடோமில் (EURATOM) 1976 வரை கேட் மாற்றப்படவில்லை; ORNLஇல் அது குறைந்தது 1979 வரை மற்றும் ஒருவேளை அதற்கு மேலோ பயன்படுத்தப்பட்டதாகத் தெரிகிறது (ஜோர்டான் மற்றும் பலர் (Jordan et al. 1976, 1977).

காட்-இன் (GAT) மூலோபாயம் நேரடியானது மற்றும் இடம்சார்ந்தது (direct and local): எளிய (wordfor-word replacement) சொல்லுக்குச்சொல் இடம்பெயர்த்தல்; அதைத் தொடர்ந்து ஒரு குறிப்பிட்ட அளவு தெளிவற்று ஆங்கிலத்தை ஒத்திருக்கிற ஒன்றை விளைவிக்கும் படி சொற்களின் இடமாற்றம் (transposition). மிக விரைவில், ஒரு "சொல்" என்பது ஒரு மரபுத்தொடரை ("idiom") உருவாக்கும் ஒற்றைச் சொல் அல்லது சொற்களின் வரிசை என வரையறுக்கப்பட்டது. காட் (GAT) வடிவமைப்பில் உண்மையான மொழியியல் கோட்பாடு எதுவும் இல்லை; மற்றும் கணினி அறிவியல்சார்ந்த (computer science) உள்ளார்ந்த கணியியல்சார் கோட்பாடு எதுவும் இல்லை. காட் (GAT) ஆனது கொடுக்கப்பட்ட உரைக்காக செயல்பட உருவாக்கப்பட்டது; பின்னர் அடுத்த உரைக்கு வேண்டி மாற்றி அமைக்கப்பட்டது; இவ்வாறு தொடரப்பட்டது. இறுதி முடிவு சிக்கலான

கண்டுபிடிக்க இயலாத ஒருசீரான ஒழுங்குமுறை (monolithic system) ஆகும்: ORNL மற்றும் யுரடோமுக்கு (EURATOM) வழங்கப்பட்ட பிறகு அது குறிப்பிடத்தக்க மாற்றம் அடையவில்லை. அது நீண்ட காலம் பயன்படுத்தப்பட்டது என்ற உண்மையில் குறிப்பிடத்தக்கது ஒன்றும் இல்லை; எந்தவொரு மாற்று மாற்றமும் இல்லாத நிலையில் மொழிபெயர்ப்பு சேவைகள் மிகவும் தேவைப்படும் பயனர்களால் குறைந்த தரமான இயந்திர மொழிபெயர்ப்பைக் கூட. பொறுத்துக்கொள்ள முடியும் என்பது இதிலிருந்து கிடைக்கும் பாடம் ஆகும்.

ஜார்ஜ்டவுணின் இயந்திர மொழிபெயர்ப்பு ஆய்வு 1960களின் மத்தியில் நிறுத்தப்பட்டது. காட் (GAT) உருவாக்கியவர்களில் ஒருவரான பீட்டர் டொமா (Peter Toma) லாட்செக்-ஐ (LATSEC) உட்படுத்தி சிஸ்ட்ரான் (SYSTRAN) ஒழுங்கமைப்பை உருவாக்கினார். இது 1970-இல் ரைட் பேட்டர்சன் ஏ.எஃப்.பி.யில் (Wright Patterson AFB) யு.எஸ்.ஏ.எஃப் வெளிநாட்டு தொழில்நுட்ப பிரிவில் (எஃப்.டி.டி.) (USAF Foreign Technology Division FTD)) ஐபிஎம் மார்க் II ஒழுங்குமுறையையும் (IBM Mark II system) 1976-இல் யுடொட்ரோமில் (EURATOM) காட்-ஐயும் இடம்பெயர்த்தது. தகவல் கையகப்படுத்தல் நோக்கங்களுக்காக ரஷ்ய மொழியை ஆங்கிலத்தில் மொழிபெயர்க்க சிஸ்ட்ரான் இன்னும் அங்கு பயன்படுத்தப்படுகிறது.

3.3. செட்டா : தானியங்கு மொழிபெயர்ப்புக்கு மையஆய்வு (CETA: CENTRE D'ÉTUDES POUR LA TRADUCTION AUTOMATIQUE)

ஸொல்கம் (Solcum1985) தமது “A survey of machine translation: its history, current status, and future prospects” என்ற கட்டுரையில் செட்டா (CETA) என்பது குறித்து எழுதிய தகவலின் படி இப்பகுதி அமைகின்றது.

1961-இல் பிரான்சில் க்ரெனொபிள் பல்கலைக்கழகத்தில் (Grenoble University) ருஷ்ய மொழியிலிருந்து பிரஞ்சுமொழிக்கு மொழிபெயர்க்கும் ஆய்வு தொடங்கப்பட்டது. காட் (GAT) போல் அல்லாமல் செட்டா (CETA) ஆய்வு தெளிவான மொழியியல் கோட்பாட்டுடன் தொடங்கப்பட்டது. குறிப்பாக ஒரு எல்லைக்குட்பட்ட சொல் இடமாற்றத்தைக் கட்டுப்படுத்த (the local approach) வாக்கியங்களுக்கிடையிலான பட்டறிவைச் சாராமல் ஒவ்வொரு வாக்கியத்தின் (மொழிச் சுதந்திரமான, சார்பற்ற உருப்படுத்ததை குறிப்பாலுணர்த்தும்) சார்பு-அமைப்பை (dependency-structure) பெற (a global approach) முடிவுசெய்யப்பட்டது. CETA-இன்

கோட்பாடு அடிப்படை இலக்கண நிலையில் இடைமொழி அணுகுமுறையாக (intelingual approach) அமைந்தது; ஆனால் சொல் (அகராதி) நிலையில் மாற்றமாக (ஒரு குறிப்பிட்ட மொழியின் பொருண்மை உருப்படுத்தத்தை மற்றொரு மொழிக்குப் பொருத்துவதைக் குறிப்பாலுணர்த்துவதாக) அமைந்தது. கணினி அறிவியலின் நிலைப்பாடு பழமையானதாக இருந்தால் க்ரொனோபிள் (Grenoble) ஐபிஎம் (IBM) கட்டக மொழியை CETA-இன் அடிப்படை மென்பொருளாகப் பின்பற்றும்படியான கட்டாயம் ஏற்பட்டது (Hutchins 1978).

செட்டா (CETA) பத்து ஆண்டுகளாக உருவாக்கப்பட்டது. 1967-71 காலகட்டத்தில் இது ரஷ்ய மொழியின் கணக்கு மற்றும் இயற்பியல் பனுவல்களின் 400,000 சொற்களை பிரஞ்சுமொழியில் மொழிபெயர்க்கப் பயன்படுத்தப்பட்டது. இக்காலகட்டத்தில் முக்கியமான கண்டுபிடிப்பு இடைமொழியைப் பயன்படுத்துவது மொழிபெயர்ப்பை எவ்வாறு வெளிப்படுத்துவது என்பதன் எல்லா தடையங்களையும் அழிப்பதாக அமையும் என்பதாகும். ஒரு செயல்பாட்டு ஒழுங்குமுறையில் எவ்வாறு மொழிபெயர்ப்பை உருவக்குவது என்பது குறித்த புறவயமான தடையங்களைத் தக்கவைப்பது மிக முக்கியமானதாகும் என்பதையும் "தோல்வி-மென்மை" ("fail-soft") நடவடிக்கைகளை ஒழுங்குமுறையில் உட்படுத்தவேண்டும் என்பதையும் செட்டா (CETA) ஆய்வாளர்கள் கற்றுக்கொண்டனர். (எடுத்துக்காட்டாக இந்தோஐரோப்பிய மொழிகள் இன உறவுள்ள சொற்களுடன் பல அமைப்பு அடிப்படையிலான ஒற்றுமையை கொண்டிருப்பதைப் பயன்படுத்தவேண்டும்). இடைமொழி இதை எளிதில் அனுமதிக்காது. 1971-இல் வன்பொருள் மாற்றம் செட்டாவைக் (CETA) கைவிடுவதைத் தூண்டியது. இது கெட்டா (GETA) என்ற புதிய ஆய்வின் அல்லது ஒழுங்குமுறையின் உய்வாக்கத்திற்குத் தொடக்கமிட்டது. இது செட்டாவின் (CETA) தோல்வி-மென்மை மாற்ற அணுகுமுறையை (fail-soft transfer design) அடிப்படையாகக் கொண்டு அமைந்தது. இந்த மென்பொருள் சொல்லத்தக்க அளவுக்கு கட்டக மொழியை (assembly language) உட்படுத்தியுள்ளது; தொடர்ந்து கட்டக மொழியைப் பயன்படுத்துவது மோசமான விளைவை ஏற்படுத்தியது.

3.4. மாண்ட்ரியல் பல்கலைக்கழத்தின் டாவம் ஆய்வு (TAUM: TRADUCTION AUTOMATIQUE DE L'UNIVERSITÉ DE MONTRÉAL)

TAUM என்ற ஆய்வுத்திட்டம் 1965-இல் மாண்ட்ரியல் பல்கலைக்கழகத்தில் கன்னடா அரசாங்க நிதியுதவியுடன் நிறுவப்பட்டது (Slocum, 1985). முழுவதும் மாற்ற அணுகுமுறைப் (transfer approach) பின்பற்றி வடிவமைக்கப்பட்ட முதல் இயந்திரமொழிபெயர்ப்பு இது என்று கூறவியலும். TAUM 6600 சிடிசியில் (பின்னர் சைபர் 173இல்) ஃபோர்ட்ரான் நிரலாக்க மொழியை (programming language) தேர்ந்தெடுத்தது. கனடா வானிலை ஆய்வு மையத்தின் (Canadian Meteorological Center) தூண்டுதலால் இவ்வாய்வுத்திட்டம் TAUM-METEO என முன்னேறியது. வானிலை முன்னறிவிப்புகள் ஓரளவு உறுதியான நடையில் இருந்தபடியாலும் ஆங்கிலத்திலிருந்து பிரஞ்சு மொழிக்கு மொழிபெயர்ப்பது கடினமானதாக இருந்தபடியாலும் இவ்வாய்வு முக்கியத்துவம் பெற்றது. 1975-இல் TAUM வானிலை முன்னறிவிப்புகளை ஆங்கிலத்திலிருந்து பிரஞ்சு மொழிக்கு மொழிபெயர்ப்பு செய்ய நியமிக்கப்பட்டது. 1977-இல் METEO இயந்திர மொழிபெயப்புக்காக நிறுவப்பட்டது.

இதைத் தொடர்ந்து TAUM ஆங்கிலத்திலிருந்து பிரஞ்சுமொழிக்கு 90 மில்லியன் சொற்கள்கொண்ட விமானப்போக்குவரத்து பராமரிப்பு கையேடுகளை (aviation maintenance manuals) மொழிபெயர்க்க எதிர்பார்க்கப்பட்டது. இதிலிருந்து TAUM பிரத்தியேகமாக விமானப்போக்குவரத்து பராமரிப்புக் கையேடுகளை மொழிபெயர்ப்பதில் கவனக்குவிப்பு செய்தது. கையேடுகளிலில் இருந்த பன்முகப் பெயர்ச்சொல் கூட்டுக்கள் (multiple-noun compounds) சவாலாக அமைந்தது. மொழிபெயப்புக் குழு 1977-இல் TAUM-AVIATION ஒழுங்குமுறையில் குறிப்பிடத்தக்க பொருண்மையியல் ஆய்வை உட்படுத்தியது.

1979-இல் நடந்த பரிசோதனைக்குப் பிறகு TAUM-AVIATION அதன் எதிர்பார்க்கப்பட்ட பயன்பாட்டின் உற்பத்திக்குச் சரியான நேரத்தில் தயாராகாது என்பது வெளிப்படையாகத் தெரிந்தது. கனடா அரசாங்கம் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் நிலையை மதிப்பிட தொடர்ச்சியான பரிசோதனைகளை ஏற்பாடு செய்தது. ஒவ்வொரு அகராதிப் பதிவின் விலை (மனித உழைப்பு, செலவு அடிப்படையில்) மிக அதிகமானது என்றும் ஒழுங்குமுறையின் வேலைசெய்யும் நேரச் செலவும் அதிகமானது (ஒரு சொல்லுக்கு 6 செண்ட்) என்றும் கண்டுகொள்ளப்பட்டது. குறிப்பாக மனித மொழிபெயர்ப்புடன் (சொல்லுக்கு 8 செண்ட்) ஒப்பிடுகையில் TAUM-இன் பின்திருத்தும் செலவு (சொல்லுக்கும் 10 சென்ண்ட்) மனித மொழிபெயர்ப்பின் செலவைவிட (சொல்லுக்கு 4 செண்ட்) அதிகமானது என்பது கருத்தில்

கொள்ளப்பட்டது. கன்னடா அரசாங்கத்தின் மோசமான பொருளாதார சூழல் போன்ற பல காரணங்களால் TAUM ஆய்வுத்திட்டம் 1981-இல் கைவிடப்பட்டது. சமீபத்தில் கனடாவில் இயந்திரமொழிபெயர்ப்பில் ஆர்வம் புதிப்பிக்கப்பட்டு இன்றைய முன்னேற்ற நிலைக் கணிப்புகளுக்கு (state-of-art surveys) ஆணையிடப்பட்டுள்ளது.

3.5. ஆல்ப்: தானியங்கு மொழி ஆய்வு (ALP: AUTOMATED LANGUAGE PROCESSING)

1971-இல் ப்ரிகாம் யங் பல்கலைக்கழகத்தில் (Brigham Young University (BUY)) மோர்மோன் திருச்சபை பனுவல்களை ஆங்கிலத்திலிருந்து பன்மொழிகளுக்கு மொழிபெயர்க்க ஒரு ஆய்வுத்திட்டம் நிறுவப்பட்டது (Slocum, 1985). இது பிரஞ்சு மொழியில் தொடங்கி போர்த்துகீசு, ஸ்பானிஷ் ஆகிய மொழிகளுக்கு விரிவுபடுத்தப்பட்டது. தொடக்கத்தில் இதன் நோக்கம் ஜங்ஜன் இலக்கண (Junction Grammar, Lytle et al 1975) அடிப்படையில் முழு தானியியங்கு ஒழுங்குமுறையை உருவாக்குவதாக அமைந்தது. ஆனால் 1973-இல் இயந்திர உதவியுடனான மொழிபெயர்ப்புக்கு (machine aided translation (MAT)) மாற்றப்பட்டது. இதன்படி இந்த ஒழுங்குமுறை வாக்கியங்களை முன் நிரலாக்கம் செய்யப்பட்ட மொழியியல் விதிகள் அடிப்படையில் தானாகவே ஆய முயலவில்லை; அதற்குப் பதிலாக மனிதருடன் ஊடாட்டம் செய்வதை நம்பியிருந்தது. இந்த ஊடாட்ட மொழிபெயர்ப்பு ஒழுங்குமுறை (interactive translation system) மனித உதவியுடன் வாக்கியங்களின் முழுமையான ஆய்வை செய்தது மற்றும் மனித உதவியுடன் மறைமுகமாக மொழிபெயர்த்தது.

ப்ரிகாம் யங் பல்கலைக்கழக ஆய்வுத்திட்டம் (BYU) செயல்படக்கூடிய ஒரு ஒழுங்குமுறையை உருவாக்கவில்லை. வன்பொருட்களின் செலவுகளும் மனித ஊடாட்டத்தின் அளவும் சிக்கலும் செலவுத்திறனை தடைசெய்தது. மோர்மோன் தேவாலயம் (Mormon Church) பல்கலைக்கழம் வழி ஆய்வுத்திட்டத்தை அகற்றுவதற்குத் தொடங்கியது. 1980-இல் ப்ரிகாம் யங் பல்கலைக்கழகத்தைச் சார்ந்த ஒரு குழும நிரலர்கள் வெயிண்டர் தகவல் தொடர்பு நிறுவனத்துடன் (Weidner Communications Corporation) சேர்ந்து முழு தானியங்கு நேரடியான வெயிண்டர் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை (direct Weidner MT system) உருவாக்க உதவினார்கள். அதே சமயத்தில் மீதியிருந்த ப்ரிகாம் யங் பல்கலைக்கழக ஆய்வுத்திட்ட அங்கங்கள் அங்கிருந்து விலகி தானியங்கு மொழி ஆய்வு ஒழுங்குமுறைகளை உருவாக்கினர் (Automated Language Processing Systems (ALPS)); அவர்கள் ஊடாட்ட மொழிபெயர்ப்பு

ஒழுங்குமுறை உருவாக்குவதைத் தொடர்ந்தனர். இந்த இரண்டு ஒழுங்குமுறைகளும் தற்போது தீவிரமாக வணிகம்செய்யப்படுகின்றது.

3.6. சிஸ்ட்ரான் (SYSTRAN)

சிஸ்ட்ராம் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை பிற இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளைவிட நீண்டகாலம் செயல்பாட்டுச் சேவையில் இருந்தது என்று ஹட்சின்ஸ் மற்றும் ஸொமர்ஸ் (Hutchins and Somers1992) குறிப்பிடுகின்றனர். சிஸ்ட்ரான் குறித்த தகவல் ஹட்சின்ச் மற்றும் ஸோமர்ஸ் எழுதிய நூலிருந்து என்பதிலிருந்து எடுத்தாளப்பட்டுள்ளது (Hutchins and Somers 1992: 175-1989).

SYSTRAN (சிஸ்ட்ரான்) என்பதன் விரிவு System Translation (ஒழுங்குமுறை மொழிபெயர்ப்பு) என்பதாகும். இது மாற்றம் அடிப்படையிலான தானியங்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையாகும். இதற்கான திட்டத்தை பீட்டர் டோமா என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன்முதலில் வெளிப்யிடப்பட்ட கணிப்பொறி வழி மொழிபெயர்ப்புக்கான மென்பொருளில் இது ஒன்றாகும் (Slocum, 1985). ஐபிஎம் மார்க் II (IBM Mark II System) ஒழுங்குமுறையையும் நாசா (NASA) 1974-இல் அப்பல்லோ-சோயுஸ் உடனுழைப்புடன் (Apollo-Soyuz collaboration) தொடர்புள்ள விஷயங்களை மொழிபெயர்ப்பு செய்ய சிஸ்ட்ராணைத் தேர்ந்தெடுத்தது. யூரடோம் (EURATOM) காட்-ஐ (GAT) சிஸ்டிரானால் இடம்பெயர்த்தது. மொழிபெயர்ப்பு ஒழுங்குமுறை தொடர்ந்து மேம்படுத்தப்பட்டு வந்தது. அதிக அளவிலான கட்டக வடிவைப்புக்கு (modular design) மாற்றப்பட்டது. தலைப்புகள் அடிப்படையிலான பொருள்விளக்கச் சொற்கோவைகள் (glossaries) (குறிப்பாக, பனுவலின் பாடப்பரப்புக்கு ஏற்றபடியான அகராதிகள்) பயன்படுத்தப்பட்டன. அமெரிக்க இராணுவ படையின் (US Army Force (USAF)) வெளிநாட்டு தொழில்நுட்பம் பிரிவு (Foreign Tchnology Division (FTD)) மில்லியனுக்கும் மேற்பட்ட பதிவுகளைக்கொண்டிருந்தது.

1976-இல் ஐரோப்பிய சமூகங்களின் ஆணையம் (Commission of the European Communities (CEC)) ஆங்கில-பிரஞ்சு மொழிபெயர்ப்பு ஒழுங்குமுறையை மதிப்பீடு செய்யவும் ஆற்றலுள்ள பயன்பாட்டிற்கும் வேண்டி வாங்கியது. தகவல் பேறை நோக்கமாகக் கொண்டிருந்த ஏஃப்டிடி (FTD), நாசா, யூரடாம் என்பனவற்றிற்கு மாறாக சியிசி (CEC) தகவல் பரப்புவதைக் நோக்கமாகக் கொண்டிருந்தது. ஆரம்ப கால மதிப்பீடுகள் எதிர்மறையாகவே அமைந்தது. அகராதி

பதிவுகளைக் கூட்டுவது மொழிபெயர்ப்பு ஒழுங்குமுறையை மேப்படுத்தும் என்று நம்பப்பட்டது. சிஸ்ட்ரானின் கூடுதல் பதிப்புகள் (பிரஞ்சு-ஆங்கிலம் 1978இலும் ஆங்கிலம்-இத்தாலியமொழி 1979-இலும்) வாங்கப்பட்டன

மேலும் 1976-இல் கனடா ஜெனரல் மோட்டார்ஸ் (General Motors of Canada) பலவகைப்பட்ட கையேடுகளை ஆங்கிலத்திலிருந்து பிரஞ்சுக்கு மொழிபெயர்க்க சிஸ்ட்ரானை வாங்கியது. 1981-இல் ஜிஎம்மின் ஆங்கில-பிரஞ்சு அகராதி 130,000 சொற்களாக விரிவடைந்தது. இதைத் தொடர்ந்து ஜிஎம் சிஸ்ட்ரானின் ஆங்கில-ஸ்பானிஷ் பதிப்பையும் வாங்கியது.

சிஸ்டிரானைப் பயன்படுத்துவோர் மொழிபெயர்ப்பு ஒழுங்குமுறையைக் கட்டுப்பாடற்ற பனுவல்களுக்கு பயன்படுத்தும்போது ஜெராக்ஸ் (Xerox) லாட்ஸ்டெக்குடன் (LATSEC) ஆலோசனைசெய்து ஒரு எல்லைக்குட்பட்ட உள்ளீட்டு மொழியை (பன்மொழிய திருத்தியமைக்கப்பட்ட ஆங்கிலம்) உருவாக்கினர். அதாவது ஜெராக்ஸ் ஆங்கிலத் தொழில்நுட்ட எழுத்தாளர்களை ஒரு சிறப்பான சொற்றொகையைப் பயன்படுத்தவும் கண்டிப்பான நடைக் கையேட்டைப் பின்பற்றவும் வேண்டிக்கொள்ளப்பட்டனர்.

தற்போது சிஸ்ட்ரான் CECயில் வழக்கமான மொழிபெயர்ப்புக்கும் அதைத் தொடர்ந்து ஆங்கிலம்-பிரஞ்சு, பிரஞ்சு-ஆங்கிலம், ஆங்கிலம்-இத்தாலிய இணைகளின் கிட்டத்தட்ட 1000 பக்கங்களைப் பிந்தைய திருத்தியமைக்கவும் பயன்படுத்தப்படுகின்றது. CEC சூழலின் வெற்றி காரணமாக ஆணைக்குழு ஆங்கில-ஜெர்மன், ஜெர்மன்-ஆங்கில பதிப்புகளுக்கு ஆணையிட்டுள்ளது.

3.6.1. வரலாற்றுப் பின்னணி

சிஸ்ட்ரான் ஒழுங்குமுறையின் தோற்றத்தை 1950களின் பிற்பகுதியில் இயந்திர மொழிபெயர்ப்பின் ஆரம்பமுயற்சிகளுக்குக் கொண்டுசெல்லலாம். சிஸ்ட்ரான் ஒழுங்குமுறையின் வடிவமைப்பாளர் பீட்டர் டோமா (Peter Toma) 1960இல் செயல்விளக்கம் அளிக்கப்பட்ட ரஷ்யன்–ஆங்கில மொழிபெயர்ப்புக்கு உரிய ஜார்ஜ்டவுண் பல்கலைக்கழக காட் ஒழுங்குமுறையின் (Georgetown University GAT system) சேர்னா (SERNA) நடைமுறைப்படுத்தத்தின் முதன்மை நிரலர் ஆவார். டொமா ஜார்ஜ்டவுண் பலகலைக்கழக திட்டத்தில் சேரும் முன்னரே கலிபோர்னிய தொழிநுட்ப நிறுவனத்தில் 1957இல் இயந்திரமொழிபெயப்பைத் தொடங்கி இருந்தார். அவர் தமக்குரிய நிறுவனத்தைத் நிறுவ 1962இல் ஜார்ஜ்டவுணை விட்டுச்சென்றார். ரஷ்யன்–ஆங்கில இயந்திர மொழிபெயர்ப்பைத்

தொடர்ந்து நடத்தி அணுசக்தி மற்றும் மருத்துவப் புலங்களின் மொழிபெயர்ப்புக்கு வேண்டி ஆட்டோட்ரான் (AUTOTRAN), டெக்னோட்ரான் (TECHNOTRAN) என்ற இரண்டு ஒழுங்குமுறைகளை உருவக்கினார். 1964-இல் டோமா ஜெர்மனிக்குச் சென்றார். அங்கு ஜெர்மன் ஆராய்ச்சி சமூகத்தின் (Deutche Forschungsgemeinschaft) ஆதரவுடன் சிஸ்ட்ரான் ரஷ்யன் ஆங்கில இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவக்கத் தொடங்கினார். இந்த நிலையில்/கட்டத்தில் ரஷ்யன் ஜெர்மன் மொழிபெயர்ப்புக்கு ஒரு ஒழுங்குமுறையை சுவீகரிக்க வேண்டி காட்-செர்னா வம்சாவளியின் தெளிவான அறிகுறிகளைக்காட்டிய மூலமுன்வகை சார்லாண்ட் பல்கலைக்கழகத்தில் மதிப்பீடு செய்யப்பட்டது. இருப்பினும் இறுதியாக சார்ப்ருக்கேன் குழு (Sarbrücken group) தங்களுடைய சொந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையான சூசியை (SUSY) உருவாக்கத் தீர்மானித்தனர்.

1968இல் டோமா அமெரிக்க விமானப்படைக்காக (United States Air Force (USAF) ரஷ்ய-ஆங்கில இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்க லா ஜோலா (La Jolla), கலிபோர்னியாவில் (California) லாட்செக் இங் (Latsec Inc.) என்ற நிறுவனத்தை நிறுவினார். ['System translation' என்பதன் தலைப்பெழுத்தன் (acronym)-ஆன] சிஸ்ட்ரான் (SYSTRON) 1969இன் தொடக்கத்தில் ரைட் பேட்டர்சன் விமானப்படைத் தளத்தில் (Wright-Patterson Air Force Base) ((டேடொன், ஓஹியோ (Dayton,Ohio)) பரிசோதனை செய்யப்பட்டது; மற்றும் ஜூலை 1970இலிருந்து இப்போது வரை USAF-இன் வெளிநாட்டுத் தொழில்நுட்பப் பிரிவில் (Foreign Technology Division) மொழிபெயர்ப்பைத் தொடர்ந்து தந்துகொண்டிருக்கின்றது. பின்னர் சிஸ்ட்ரான் நாசாவால் (NASA) யுஎஸ்-யுஎஸ்எஸ்ஆர் (US-USSR) கூட்டு அப்பல்லோ-சோயுஸ் (Apollo-Soyuz) விண்வெளித் திட்டதின் (1974-75) போது பயன்படுத்தப்பட்டது; 1976-இல் யுரடோமில் (Euratom) ஜார்ஜ்டவுன் மொழிபெயர்ப்பு ஒழுங்குமுறையால் இடம்பெயர்க்கப்பட்டது.

மிகவும் இன்றியமையாத வளர்ச்சி, ஜூன் 1975இல் ஐரோப்பிய சமூகங்களின் ஆணையத்தின் (Commission of the European Communities (CEC)) பிரதிநிதிகளுக்கு சிஸ்டிரானின் முன்மதிரி ஆங்கில-பிரஞ்சு பதிப்பின் செயல்விளக்கம் ஆகும்; இதன் விளைவாக ஐரோப்பிய சமூகங்களின் மொழிகளுக்கு இடையில் மொழிபெயர்ப்பின் பதிப்புகளை உருவக்க ஒரு ஒப்பத்தம் முடிக்கப்பட்டது. டோமாவின் உலக மொழிபெயர்ப்பு மையம் (World Translation

Centre (WTC)) என்ற புதிய நிறுவனத்துடனான ஒப்பந்தம், ஐரோப்பிய சமூகங்களின் ஆணையத்தின் மொழிபெயர்ப்புத் துறையின் ஊழியர்களால் ஒழுங்குமுறைகளின் கணிசமான வளர்ச்சிக்கான (உருவாக்கத்திற்கான) உடன்படிக்கையையும் உட்படுத்தும். செயலாக்கம் 1976 பெப்ரவரியில் ஆங்கில-பிரஞ்சு பதிப்புக்காகத் தொடங்கப்பட்டது; இது பிரஞ்சு-ஆங்கிலம், ஆங்கிலம்-இத்தாலிமொழி இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் உருவாக்கத்தால் தொடரப்பட்டது. 1981 மார்ச்சில் ஒவ்வொரு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையும் லக்சம்பர்க்கில் (Luxembourg) முன்னோடி உற்பத்தி சேவையை (pilot production service) நிர்மாணிக்க நியாயமான போதுமான வெளியீட்டை உற்பத்தி செய்தது. அன்றிலிருந்து CEC பிற பல மொழி இணைகளுக்கு செயல்பாட்டுப் பயன்பாட்டில் உள்ள சிஸ்ட்ரான் மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்கியது.

மேலும் அமெரிக்காவில் சிஸ்ட்ரானை ஊக்குவிக்கவும் முன்னேற்றவும் பல நிறுவனங்கள் நிறுவப்பட்டன; எ.கா. ஜெர்மனியில் சிஸ்ட்ரான் நிறுவனம் (Systran Institute in Germany) , கனடாவில் உலக மொழிபெயர்ப்பு நிறுவனம் (World Translation Corporation in Canada), ஜப்பானில் சிஸ்ட்ரான் நிறுவனம் (Systran Corporation of Japan). பிந்தையது 1980களில் ஜப்பான்மொழி-ஆங்கிலம் மற்றும் ஆங்கிலம்-ஜப்பான்மொழி மொழிபெயர்ப்புக்கு ஒழுங்குமுறைகளை உருவாக்கியது. மேலும் பிற நிறுவங்களுடன், குறிப்பாக சொல்சார் தரவுத்தளங்களின் கூட்டு உருவாக்கத்திற்கு பயனர்களுடன் ஒப்பந்தங்கள் செய்யப்பட்டன. பல ஆண்டுகளாக நீண்ட தொடர்ச்சியான பேச்சுவார்த்தைகளுக்குப் பின்னர் பிரான்சிலுள்ள கச்சோட் நிறுவனம் (Gachot company) மொழிபெயர்ப்பில் ஈடுபட்ட அமெரிக்க மற்றும் ஐரோப்பிய நிறுவங்களை வாங்கியதோடு கலவைத்தன்மையான சூழல் எளிதாக்கப்பட்டது. 1986 இலிருந்து வெளியில் இருக்கிற சிஸ்ட்ரானின் உரிமையை வைத்திருக்கிற ஒரே நிறுவனம் ஜப்பானின் சிஸ்ட்ரான் நிறுவனத்தை சொந்தமாகக் கொண்ட IONA நிறிவனம் ஆகும்.

இப்போது சிஸ்டிரானைப் பயன்படுத்துபவர் பலர் உள்ளனர் மற்றும் மொழி இணைகளின் எண்ணிக்கையும் ஆண்டுக்கு ஆண்டு வளர்ந்தது. கீழ்வரும் அட்டவணை இருக்கிற மொழிச் சேர்க்கைகள் மற்றும் வளர்ச்சி இவற்றைக் காட்டுகின்றன. மேலும் ஸொய்சி-ஸொவ்ஸ்-மாண்ட்மோரென்சி-யில் (Soisy-sous-Montmorency (பாரிஸ் பக்கத்தில்) செண்ட்ரல் கணினியிலிருந்து கசோட் நிறுவனம் பல இணைகளுக்கு மொழிபெயர்ப்பு சேவைகளைத்

தருகின்றது; சில பிரான்சில் மினிடெல் நெட்வொர்கில் (Minitel network) இருக்கின்றது: பொதுமக்களின் பயன்பாட்டிற்கு முதல் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை.

சிஸ்ட்ரானின் முதிர்ச்சி சமிக்ஞை லக்செம்பர்க்கில் நடைபெற்ற ஐரோப்பிய சமூகங்களின் ஆணையத்தால் (Commission of the European Communities) ஏற்பாடு செய்யப்பட்ட உலக சிஸ்ட்ரான் மாநாடு (World Systran Conferene) ஆகும். இந்த மாநாடு அனுபவங்களையும் பரிமாறிக்கொள்ளவும் எதிர்கால முன்னேற்றங்களைப் பற்றி விவாதிக்கவும் எல்லா முக்கிய சிஸ்ட்ரான் பயன்பாட்டாளர்களையும் ஒன்றுசேர்த்தது.

சிஸ்ரான் மொழி இணைகள் (Hutchins and Somers, 1992: 177)

உள்ளவை	உருவாக்கப்படுவை
ஆங்கிலம்⇔ பிரஞ்சுமொழி	ஆங்கிலம் ⇔சைனீஸ் மொழி
ஆங்கிலம் ⇔ஜெர்மன்மொழி	ஆங்கிலம்⇔ கொரிய மொழி
ஆங்கிலம்⇔ ஜப்பனிய மொழி	ஆங்கியல் →அரபிய மொழி
ஆங்கிலம்⇔ ரஷ்யமொழி	ஆங்கிலம் →டானிஷ் மொழி
ஆங்கிலம் ⇔ஸ்பானிஷ்	ஆங்கிலம்→ டச்மொழி
ஆங்கிலம் ⇔இத்தாலிய மொழி	ஆங்கிலம்→ ஃபினிஷ் மொழி
ஆங்கிலம் ⇔போத்துக்கீசிய மொழி	ஆங்கிலம் →நார்வேஜியன் மொழி
ஜெர்மன்மொழி ⇔பிரஞ்சுமொழி	ஆங்கிலம் →ஸ்வெதிஷ் மொழி
ஜெர்மன்மொழி⇔ இத்தாலியமொழி	பிரஞ்சு →டச் மொழி
ஜெர்மன்மொழி ⇔ஸ்பானிஷ் மொழி	பிரஞ்சு →ஜெர்மன் மொழி
	பிரஞ்சு →இத்தாலிய மொழி
	இத்தாலியமொழி →ஆங்கில மொழி
	போர்துகீசிய மொழி→ ஆங்கிலம்

3.6.2. அடிப்படை ஒழுங்குமுறை

சிஸ்ட்ரான் மேம்பாட்டாளர்களின் மொழி இணைகளின் பரந்த பரப்பெல்லையை உருவாக்கும் திறமை, இந்த ஒழுங்குமுறையின் வடிவமைப்பில் ஒப்பீட்டளவில் அதிக அளவிலான

கூறுநிலைமையை (modularity) சுட்டிக்காட்டுகின்றது. இது பின்வரும் பண்புக்கூறுகளில் பிரதிபலிக்கின்றது. இரண்டு முக்கியமான நிரல்களின் வகைகள் இருக்கின்றன: (அ) அசம்பிளர் குறியத்தில் எழுதப்பட்ட ஒழுங்குமுறை நிரல்கள்; இவை குறிப்பிட்ட மொழிகளிலிருந்து சுதந்திரமாக இருக்கின்றது. இவை கட்டுப்பாடு (control) மற்றும் பயன்பாடு (utility) நிரல்கள் ஆகும்; இவை அகராதி தேடல் நடைமுறைகளுக்குப் பொறுப்பாகும் (ஆ) மொழிபெயர்ப்பு நிரல்கள் பல நிலைகளாகப் பிரிக்கப்பட்டுள்ளன; ஒவ்வொன்றும் தனி நிரல் தொகுதியைக் கொண்டுள்ளன. பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் உரிய மொழிபெயர்ப்பு நிரல்கள் குறிப்பிட்ட மொழி இணைகளிலிருந்து சுதந்திரமானவை என்று கூறப்படுகின்றது; எந்த இலக்கு மொழி சம்பந்தப்பட்டாலும் ஒரு குறிப்பிட்ட மொழிக்கான பகுப்பாய்வுத் தொகுதி நிலையானது. மேலும் ஒரு பொதுவான ரொமான்ஸ் மொழி (Romance language) பகுப்பாய்வு 'ட்ரங்' (trunk) உருவாக்கப்பட்டுள்ளது; எந்த ரொமான்ஸ் மொழி மூலமொழியானாலும் (பிரஞ்சு, இத்தாலிய மொழி, ஸ்பானிஷ் மொழி, போர்த்துகீசிய மொழி) இதைச் செயல்படுத்த இயலும். காசோட் (Gachot) சிஸ்டிரானை பெற்றபின்னர் பெரும்பலான இந்தத் கூறுநிலைமை நிறைவேற்றப்பட்டது. புதிய உத்திகள் எப்பொதெல்லாம் தகுந்ததாகத் தோன்றுகின்றதோ அவற்றை ஒப்பீட்டளவில் நேரடியாகப் புகுத்த இது இயலச்செய்கின்றது.

இருப்பினும் சில அம்சங்களில் சிஸ்டிரானின் அடிப்படை செயல்முறைகள் USAF ரஷ்யன்-ஆங்கில ஒழுங்குமுறைக்கு முதலில் கருதப்பட்டது போல அவ்வாறே இருக்கின்றது. முக்கியமான கூறு பெரிய இருமொழிய அகராதிகளாக இருக்கின்றன; அவை பகுப்பாவின் போதும் உருவாக்கத்தின் போதும் பயன்படுத்தப்படும் சொல் நிகரன்களையும் (lexical equivalents) இலக்கண மற்றும் பொருண்மையியல் தகவல்களையும் கொண்டிருக்கின்றன. பெரும்பாலான இந்தத் தகவல்கள் வழிமுறைகள் (algorithms) வடிவில் இருக்கின்றன; அவை மொழிபெயர்ப்பு செயலாக்கத்தின் பல நிலைகளின் போது செயல்படுத்தப்படும். அமைப்புப் பகுப்பாய்வின் நிரல்கள் மற்றும் உருவாக்கம் பெரும்பாலும் சுதந்திரமாக இருக்கிறது; ஆனால் முக்கியமான மொழிபெயர்ப்பு செயற்பாங்குகள் கணிசமான சிக்கலான இருமொழிய அகராதிகளால் இயக்கப்படுகின்றது. இருப்பினும் சிஸ்டிரானின் புதிய பதிவுகளுக்கான இருமொழிய அகராதிகளின் தொகுப்பு எப்போதும் தொடக்கநிலையிலிருந்து தொடங்கப்பட

வேண்டியதில்லை; பல நடவடிக்கைகளில் உள்ளது போன்று புதிய இலக்குமொழிகளுடன் இணைக்கும் போது மூலமொழிச் சொற்களின் குறியமாக்கத்தில் சிறிய வேறுபாடுகள் செய்தால் போதுமனது. இந்த உண்மை அண்மையில் பல வேறுபட்ட இலக்கு மொழிகளுக்கு (பல்-இலக்கு (multi-target)) நிகரன்களைக் கொண்ட அகராதிகளைத் தொகுப்பதை ஊக்குவித்தது.

3.6.2.1. அகராதிகள்

சிஸ்டிரானின் தரவுத்தளங்கள் முக்கியப் பகுதி அகராதி (main stem dictionary), தனிச் சொல் பதிவுகளின் இருமொழிய அகராதி (bilingual dictionary of single-word entires), பல பல்சொல் 'சூழல்சார்' அகராதிகள் (multi-word 'contextual' dictionaries) எனப் பகுக்கப்பட்டுள்ளன. முக்கியப் பகுதி அகராதியில் ஒவ்வொரு மூல மொழிச் சொல்லும் (அதன் வேர் வடிவில்; முழுவடிவுகள் தரப்பட்டுள்ள ஆங்கிலம் தவிர) முழுமையான உருபனியல், தொடரியல், பொருண்மையியல் விளக்கம் தரப்பட்டுள்ளன: இலக்கண வகைப்பாடு, சொல் இயைபு (government), இணைதிறன், உடன்பாடு, செயப்படுபொருண்மை (transitivity), பெயர் வகை ('விலங்கு', 'எண்ணக்கூடியவை', 'அருவம்', போன்றன), பொருண்மைக் குறிப்பான்கள் ('பௌதிகப்பண்பு', 'கொள்வான்'. 'கருவி', 'உணவுப்பொருள்' போன்றன); மேலும் அடிச்சொல் வடிவை நிகரான இலக்குச் சொல்லாக மொழிபெயர்த்தல்; இது உருவாக்கத்திற்குத் தேவையான தகவல்களையும் கொண்டிருக்கும். தனித்தனியான பதிவுகளைக் கொண்டிருக்கும் வேறு இலக்கண வகைபாடுகள் கொண்ட ஒப்புருச்சொற்களுக்கும் (homographs) ஒரே வகைப்பாட்டைச் சார்ந்த ஒப்புருச் சொற்களுக்கும் இடையில் வேறுபாடு செய்யப்பட்டிருக்கும்; அவை 'சூழல் அகராதிகளால்' பல்பொருளொருமொழிகள் போன்று (polysemes) கையாளப்படும். இந்த வேறுபாடு மொழிபெயர்ப்புக்குத் சிஸ்ட்ரானின் பெரும்பான்மைத் தொடரியல் அடிப்படையான அணுகுமுறையின் பிரதிபலிப்பாகும்; அதாவது முதலில் தொடரியல் சிக்கல்கள் கையாளப்படும்; பின்னர் மீதி சிக்கல்களைத் தீர்ப்பதற்குப் பொருண்மையில் தகவல்கள் செயலாக்கப்படும். ஒவ்வொரு மூலமொழி பதிவுக்கும் ஒரேயொரு இலக்குமொழி நிகரனே தரப்பட்டுள்ளது என்பதைக் கவனத்தில் கொள்ளவேண்டும்: இதன் விளைவு இயல்புநிலை மொழிபெயர்ப்பாகும் ('default' translation); பிற அகராதிகள் இதை மாற்றாவிடில் இது அவ்வாறே இருக்கும்; எடுத்துக்காட்டாக station என்ற ஆங்கிலச் சொல்லுக்கு poste இயல்புநிலை பிரஞ்சு மொழிபெயர்ப்பாகும்.

சிஸ்டிரானின் 'சூழல்' அகராதிகள் (Systran 'contextual' dictionaries) தானியக்கமாக ஒரு தனி மூலமொழி அகராதியிலிருந்து ஆக்கப்படும்; இது ஒழுங்காக மேம்படுத்தப்படும். அவை சூழல் அடிப்படையில் பகுப்பாய்வையோ மொழிபெயர்ப்பையோ மாற்றுவதைச் சாத்தியமாக்கவும் பகுப்பாய்வு, மொழிபெயர்ப்பு இவற்றின் பல்வேறு நிலைகளில் பயன்படுத்தப்படும் ஒரு கூட்டம் அகராதிகளை உருவாக்கவும் தரவைத் தரும்.

அ) 'மரபுச்சொல்' அகராதி ('idiom' dictionary) மாறாத (நிலையான) வெளிப்பாடுகளைக் (எ.கா. on the one hand, in order to) கையாளுவதற்கு வடிவாக்கம் செய்யப்பட்டுள்ளது; இவை சில நேர்வுகளில் ஒரு தனி இலக்குமொழி வடிவை ஒத்திருக்கும்.

ஆ) 'எல்லைக்குட்பட்ட பொருண்மையியல்' அகராதி ("limited semantic' dictionary) பெயர்த்தொடர்களுக்குள் தொடரியல் உறவுகளின் இலக்கை விளக்கும், எ.கா. hydraulic brake என்பதை ஒரு சொல்சார் அலகாக அடையாளங்கண்டு கொண்டு hydraulic brake fluid என்பதை hydraulic என்பது fluid என்பதை விசேடணம் செய்யும் பகுப்பாய்வைத் தடை செய்யும். சீரான மொழிபெயர்ப்பை உறுதி செய்ய பிற கூட்டுக்கள் சொல்சார் அலகுகளாக அடையாளங்காணப்படலாம் (எ.கா. machine translation என்பதை *transduction de machine* என்று மொழிபெயர்க்காமல் *traduction automatique* என்று மொழிபெயர்ப்பது). எனவே இது ஒரு தனிப் பொருண்மை அலகாக வரும் பெயர்த்தொடர்களின் பதிவுகளையும் உள்ளடக்கும், எ.கா. (potato என்ற தனி ஆங்கிலச்சொல்லுக்கு ஒப்பாக வரும்) பிரஞ்சுமொழியின் *pomme de terre*. சில சிஸ்ட்ரானின் பதிவுகளில் இந்த இரு செயற்பாடுகளும் வேறுபட்ட அகராதிகளுக்கு இடையில் பகுக்கப்பட்டிருக்கும். 'எல்லைக்குட்பட்ட பொருண்மையியல்' அகராதி பொருண்மை மயக்கமுள்ள கோர்வைகளைப் பெயர்த்தொடர்களாக பொருள்கோள் செய்வதை உறுதி செய்யப் பயன்படுத்தப்படுகின்றது, எ.கா. equipment cooling என்பதை 'equipment which is cooling' என்பதற்குப் பதிலாக 'cooling of equipment' என்று பொருள்கொள்வது.

இ) ஒப்புருச்சொல் அகராதி ('homograph' dictionary) சில ஒப்புருசொற்களின் தீர்வுக்காக தொடரியல் சூழல்சார் தகவல்களைப் பட்டியலிடுகின்றது. எடுத்துக்காட்டாக, பிரஞ்சுமொழியில் செயப்படுபொருள் குன்றா வினைக்கும் (எ.கா. *pendre*) அதன் நேரடியான பெயர் செயப்படுபொருளுக்கும் (எ.கா. chapeau) இடையில் பொதுவாக ஒரு வரையறு அடை (எ.கா. சார்படை *un* அல்லது உடைமை மாற்றுப் பெயர் *son*) இருக்கும்; ஆனால் சில விதிவிலக்குகள்

உண்டு (எ.கா. *pendre note* 'make a note of'); அவை ஒப்புருச்சொல் அகராதியில் குறிப்பிடப்பட்டிருக்கும்.

ஈ) 'பகுப்பாய்வு' அகராதிகள் ('analytic' distionaries) குறிப்பிட்ட சொற்களுக்குப் பயன்படுத்தப்படும் பொதுவான தொடரியல் விதிகளுக்கு விதிவிலக்குகளைக் கொண்டிருக்கும். எடுத்துக்காட்டாக *nor* என்பது இணைப்புகிளவிகளுக்கான 'இயல்பான' விதியைப் பிரிக்கும்; இதில் இதைத் தலைகீழான எழுவாய் பெயர் (inverted subject noun) மற்றும் வினையால் பின் தொடர இயலும், எ.கா. …nor could he see the difficulties. இந்த அகராதிகள் பகுப்பாய்வின் பல்வேறு நிலைகளில் இயங்கும்.

உ. கட்டுப்பாட்டுப் பொருண்மையியல்சார் அகராதி (conditional semantic dictionary) இறுதி இலக்கு மொழிச் சொல் தெரிவைச் செய்ய மாற்றல் நிலையில் (state of transfer) தலையிடும். இது சாத்தியமான நிகரன்களுக்கு இடையில் உள்ள வேறுபாட்டை அறியத் தொடரியல் மற்றும் பொருண்மையியல் தகவல்களை இணைத்துக்கொள்கின்றது. எடுத்துக்காட்டாக grow என்ற ஆங்கிலச் சொல்லின் இயல்புநிலை மொழிபெயர்ப்பு பிரஞ்சு மொழியில் *grandir* ஆகும்; ஆனால் ஒரு 'விலங்கு' நிரப்பியுடன் இது *élever* ஆகும் மற்றும் 'plant'ஐ செய்படுபொருளாகக் கொண்டு இது *cultiver* ஆகும். சில நேர்வுகளில் சூழல் குறிப்புகள் அதிகமாக இருக்கும்: oil என்ற ஆங்கிலச் சொல்லுக்கு வேண்டி 400 பதிவுகள் *huile* மற்றும் *pétrole* என்ற மொழிபெயர்ப்புகளை வேறுபடுத்துகின்றது.

3.6.2.2. கணினிசார் பண்புக்கூறுகள்

முன்னர் கூறியபடி அடிப்படையில் சிஸ்ட்ரானில் இரண்டு வகை நிரல்கள் உள்ளன. அசம்பிளர் குறியத்தில் எழுதப்பட்ட 'ஒழுங்குமுறைகள் நிரல்கள்' ஈடுபடுத்தப்பட்ட மொழிகளிலிருந்தும் சுதந்திரமானதாகும்; மற்றும் இது உள்ளீடு, அகராதி தேடல் செயல்முறைகள் மற்றும் மொழிபெயர்ப்பு செயற்பாங்குகளின் (translation கட்டுப்பாடுகளுக்கு வேண்டி பொதுவான நிரல்களையும் உள்ளடக்கும். 'மொழிபெயர்ப்பு' நிரல்கள் கையாளப்பட்ட குறிப்பிட்ட மொழிகளின் படி வேறுபடும். இவைகள் மூலமொழியின் பகுப்பாய்வுக்கு வேண்டியும் மாற்றலுக்கு வேண்டியும் உருவாக்கத்திற்கு வேண்டியும் பிரிக்கப்பட்டுள்ளன. இந்த நிரல்கள் உயர்நிலைப் 'பெருமொழியில்' (higher level macro-language) எழுத்தப்பட்டுள்ளன.

சிஸ்ட்ரான் ஒரு வாக்கியத்தில் உள்ள ஒரு சொல்லுக்கு ஒன்று எனப் பதிவுகளின் வரிசைகளைக் (sequence of records) கொண்ட ('பைட் பரப்பளவுகள்' (bite areas)) நேரியல் தரவு அமைப்பைப் (linear data structure) பயன்படுத்துகின்றது. ஒவ்வொரு பைட் பரப்பளவும் சொல் மற்றும் இலக்கணத் தகவல்கள் மற்றும் அகராதி பதிவுகளில் சொல்லுடன் தொடர்புடைய மொழிபெயர்ப்பு நிகரன்கள் என்பனவற்றைக் கொண்டிருக்கும். மரபுப்படி ஒவ்வொரு பைட்டும் /இருமஎண்ணும்) தகவலின் ஒரு வகையை சேமித்துவைக்கும். எடுத்துக்காட்டாக பைட்/இருமஎண் 1 சொல்லின் முதன்மை வகைப்பாட்டை (வினை, பெயர், வினையடை) காட்டும்; பைட்/இருமஎண் 2 வினையின் மூவிடப்பெயர், மற்றும் எண் இவற்றைக் காட்டும்; பைட்/இருமஎண் 3 பெயரின் (எழுவாய் வேறுமை, செயப்படுபொருள் வேற்றுமை போன்ற) வேற்றுமையையோ வினையின் காலம், வினைநோக்கு, வினைப்பாட்டு என்பனவற்றையோ காடும்; பைட்/இருமஎண் 4 பெயரின் பால், எண் இவற்றைக் காடும்; இவ்வாறு ஏனையவற்றைக் காட்டும். குறிப்பிட்ட பைட்டின்/இருமஎண்ணின் அடைகள் பிற பைட்டுகளின்/இருமஎண்களின் மதிப்புகளுடன் ஒப்பீட்டளவில் மாறும் (எ.கா. பைட்/இருமஎண் 1 'பெயர்' அல்லது 'வினை' என்பதைப் பொறுத்து பைட்/இருமஎண் 3 மாறும்); ஆனால் சிஸ்டிரானின் நடைமுறைகள் நிலையான பைட்/இருமஎண் இடங்களில் சேகரிக்கப்பட்ட தகவலைச் சுட்டிக்காட்டும் என்பது கருதவேண்டிய அவசியமான பண்புக்கூறு ஆகும். 'எல்லைக்குட்பட்ட பொருண்மையியல்' வெளிப்பாட்டை கையாளும் விதி 'பெரு மொழியில்' குறியக்கம் செய்யப்பட்டிருக்கலாம்; எடுத்துக்காட்டாக current practice என்பது 'பெரு மொழியில்' குறியக்கம் செய்யப்பட்டிருக்கலாம் (Hutchins and Somers, 1992: 179).

CURRENT SC-B26 PRACTICE (PW)

பைட்/இருமஎண் 26 பெயரடைசார் அடையைக் குறிப்பிடும். இந்நேர்வில் முதன்மை சொல் (PW) *practice* என்பது *current* என்பதைச் குறிப்பிட்டுக்காட்டினால் இந்த விதி பயன்படுத்தப்படும். இது போன்று பைட்/இருமஎண் `102ஆல் சுட்டிக்காட்டப்பட்ட சொல் ATTACH என்ற 'பொருண்மைப் பண்புக்கூறைக்' கொண்டிருந்தால் பின்வரும் விதி வெல்லும்; அதாவது *remove* என்பது *clamp*, *belt* போன்ற நேரடி செயப்படுபொருளைக் கொண்டிருக்க வேண்டும்.

REMOVE SC-B102 ATTACH

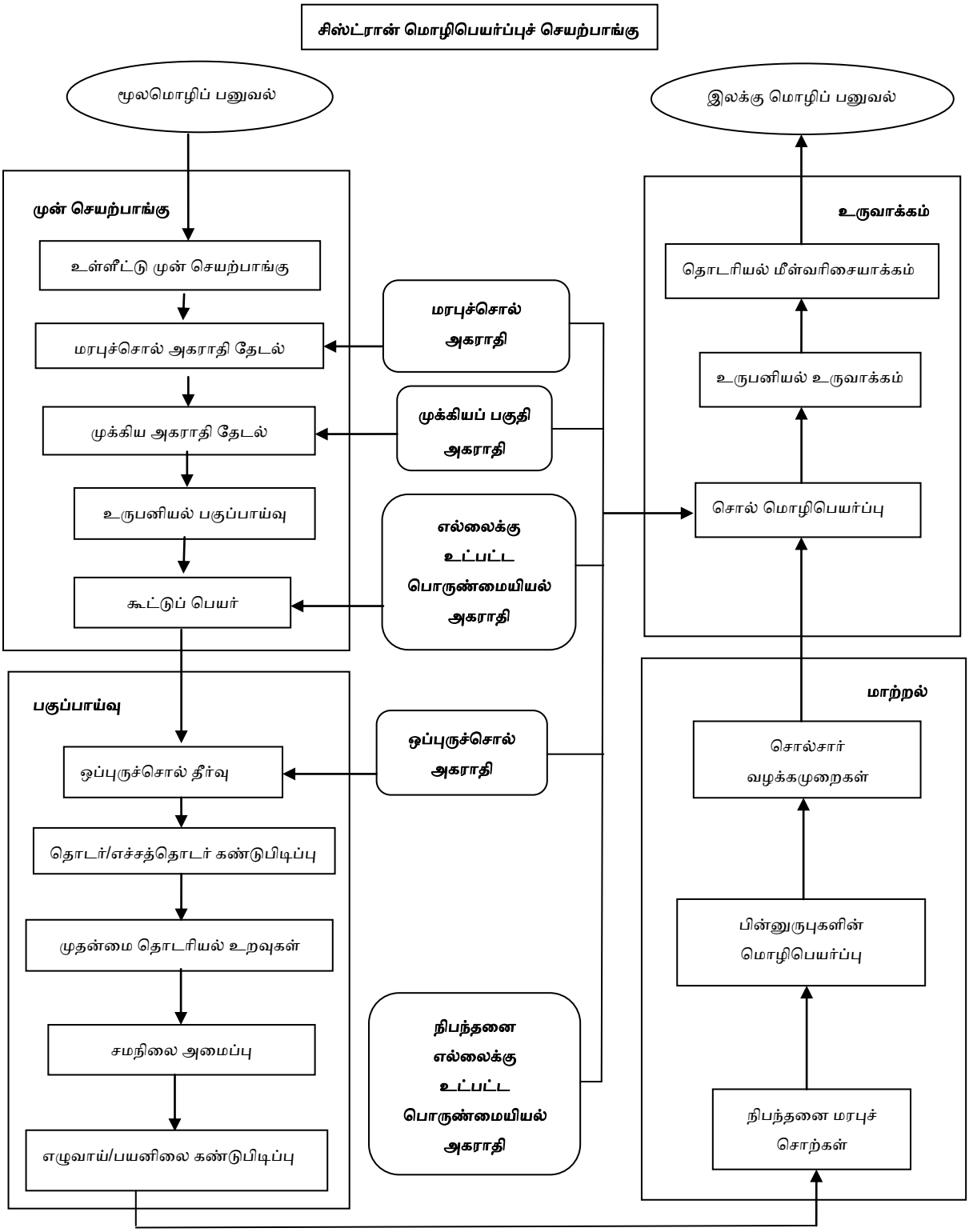
பிற பைட்கள்/இருமஎண்கள் துணை வழக்கமுறைகளின் தொடக்கத்தைத் தூண்டலாம், எ.கா. 'ஒப்புருச்சொல் அறிவிப்பு பைட்/இருமஎண்' ('homograph notification bite') ஒப்புருச்சொல் தீர்வு (homograph) நடைமுறையைத் தொடங்கும்.

தொடரியல் பகுப்பாய்வின் போது வேறுபட்ட சொற்களின் குறிப்பிட்ட பைட்களுக்கு/இருமஎண்களுக்கு இடையில் சுட்டிக்காட்டிகள் (pointers) அமைக்கப்பட்டிருக்கும், எ.கா. பெயரடைக்கும் அதை ஆளும் பெயருக்கும் (goveining noun) இடையிலுள்ள இணைப்பு, பெயரடையின் பைட்/இருமஎண் 16-இலிருந்து பெயரின் முகவரிக்கு ஒரு சுட்டிக்காட்டியாலும் மற்றும் பெயரின் பைட்/இருமஎண் 26-இலிருந்து அடை செய்யும் பெயரடையின் முகவரிக்கு ஒரு சுட்டிக்காட்டியாலும் பதிவுசெய்யப்படும். இது போன்று எழுவாய்ப் பெயர் கண்டுபிடிக்கப்பட்ட பின்னர் அதற்கும் பயனிலையின் முதற் சொல்லுக்கும் இடையில் சுட்டிக்காட்டிகள் அமைக்கப்படும்; இது கருதப்பட்ட சொற்களில் உள்ள குறிப்பிட்ட இருமஎண்களைப்/ பைட்டுகளைப் பயன்படுத்திச் செய்யப்படும்.

மேற்சொன்ன சுருக்கமான விளக்கம் காட்டுவதுபோல் சிஸ்டிடானின் தரவு அமைப்பும் கீழ்-நிலை நிரலாக்கமும் (low-leval programming) 1960களின் மற்றும் 1970களின் கணினிசார் நடைமுறைகளைத் தொடர்ந்து பிரதிபலிக்கின்றது (Hutchins and Somers, 1992: 180); இது பல 'முதல் தலைமுறை' இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் ('first generantion' MT systems) பொதுவான தன்மையாகும். இருப்பினும் அவைகள் பல அம்சங்களில் மாற்றங்களையோ மேம்பாடுகளையோ தடைசெய்யவில்லை.

3.6.2.3. மொழிபெயர்ப்புச் செயற்பாங்குகள்

மொழிபெயர்ப்பின் அடிப்படை நிலைகள்: முன்செற்பாங்கு (pre-processing), பகுப்பாய்வு (analysis), மாற்றல் (transfer), உருவாக்கம் (synthesis) என்பன (பின்வரும் படத்தில் காட்டப்பட்டுள்ளது). 'பகுப்பாய்வு', 'மாற்றல்', 'உருவாக்கம்' என்ற சொற்களின் பயன்பாடு சில விதங்களில் பிற ஒழுங்குமுறைகளில் காணப்படும் பயன்பாடுகளிலிருந்து வேறுபடும் (Hutchins and Somers, 1992: 180).


சிஸ்ட்ரான் மொழிபெயர்ப்புச் செயற்பாங்கு
மூலமொழிப் பனுவல்
இலக்கு மொழிப் பனுவல்
முன் செயற்பாங்கு
உள்ளீட்டு முன் செயற்பாங்கு
மரபுச்சொல் அகராதி தேடல்
முக்கிய அகராதி தேடல்
உருபனியல் பகுப்பாய்வு
கூட்டுப் பெயர்
மரபுச்சொல் அகராதி
முக்கியப் பகுதி அகராதி
எல்லைக்கு உட்பட்ட பொருண்மையியல் அகராதி
உருவாக்கம்
தொடரியல் மீள்வரிசையாக்கம்
உருபனியல் உருவாக்கம்
சொல் மொழிபெயர்ப்பு
பகுப்பாய்வு
ஒப்புருச்சொல் தீர்வு
தொடர்/எச்சத்தொடர் கண்டுபிடிப்பு
முதன்மை தொடரியல் உறவுகள்
சமநிலை அமைப்பு
எழுவாய்/பயனிலை கண்டுபிடிப்பு
ஒப்புருச்சொல் அகராதி
நிபந்தனை எல்லைக்கு உட்பட்ட பொருண்மையியல் அகராதி
மாற்றல்
சொல்சார் வழக்கமுறைகள்
பின்னுருபுகளின் மொழிபெயர்ப்பு
நிபந்தனை மரபுச் சொற்கள்

முதல் நிலைகள் அகராதி தேடல் (dictionary look-up) மற்றும் உருபனியல் ஆய்வு இவற்றை உள்ளடக்கும்; இவற்றை முழுப் பனுவலுக்கும் பயன்படுத்தும்.

1. உள்ளீடு (input): நிரல் பனுவலை ஏற்றும் மற்றும் வடிவாக்கத் தகவல்களை கண்டுபிடிக்கும், எ.கா. தலைப்புகள், பத்திகள், ஓர உள்தள்ளல் (indentation).

2.மரபுச்சொல் அகராதி தேடல் (idiom dictionary look-up): வேறுபடாத வடிவங்கள் (நிலையான வெளிப்பாடுகள்) கண்டுபிடிக்கப்படும்; அவற்றிற்கு தனி இலக்கண வகைப்பாடு ஒதுக்கப்படும் (எ.கா. in order to என்பதை முன்னுருபாக கொள்ளுதல்).

3. முக்கிய அகராதி தேடல்: பனுவலின் மீதி சொற்கள் தேடப்படும். இந்தத் தகவல் தரவு அமைப்பின் பைட்/இருமஎண் பரப்புகளில் நகலெடுக்கப்படும்.

4. உருபனியல் பகுப்பாய்வு: தகுதியான நேரத்தில் உருபனியல் பகுப்பாய்வு அகராதி தேடலின் போது உயிர்ப்பிக்கிப்படும். ஆங்கிலம் மூல மொழியாக இருந்தால் உருபனியல் பகுப்பாய்வுக்குத் தேவையில்லை; ஏனென்றால் அகராதிகள் முழு வடிவங்களையும் கொண்டிருக்கும். ஆனால் ரஷ்யமொழி அல்லது பிரஞ்சுமொழி போன்ற மொழிகளின் நேர்வுகளில் பகுதிகளும் விகுதி உருபுகளும் தனியாக அகராதியில் உள்ளிடப்படும். முக்கிய அகராதியில் காணப்படாத எந்த சொல்லுக்கும் இலக்கண மற்றும் வகைப்பாடுத் தகவல்களைப் ஊகிக்க இது பயன்படுத்தப்படலாம்.

5. 'எல்லைக்குட்பட்ட பொருண்மையியல்' அகராதியிலிருந்து தகவலைப் பெற்று கூட்டுசொற்கள் கண்டுபிடிக்கப்படுகின்றன; எல்லைக்குட்பட்ட அகராதியில் வரும் எல்லாச் சொல்சார் அலகுகளும் கூட்டுக்களாகவே கையாளப்படுகின்றன என்பதைக் கருத்தில் கொள்ளவும். கூட்டாக வரும் இரண்டு சொற்கள் (தனிச் சொற்களாக) வரிசையில் வர நேர்ந்தால் இது சிக்கலைத் தரும்; எடுத்துக்காட்டாக *femme de ménage* ('charlady') என்ற கூட்டு அதன் எதிர்பார்க்கப்படும் அர்த்தம் ஆ-ஆக இருக்கையில் அது அ-வில் வரும்.

அ. Il parla à la femme de ménage.

ஆ. He spoke to the woman about housekeeping

பகுப்பாய்வின் அடுத்த நிலைகள் ஒவ்வொரு வாக்கியத்திற்கும் அடுத்தடுத்து பயன்படுத்தப்படும்:

6. ஒப்புருச்சொல்லின் தீர்வு (resolution of homgraph) ஒவ்வொரு வாக்கியத்தின் தனிக் கடத்தலில் அண்மையில் வரும் சொற்களின் வகைப்பாடுகளைப் பரிசோத்தித்து பெறப்படும். இது குறிப்பாக

ஆங்கிலத்தின் நேர்வில் கலவைத்தன்மையானதாகும்; ஏனென்றால் ஒவ்வொரு சொல்லும் (அதாவது பெரும்பாலான ஆங்கிலச் சொற்கள்) சாத்தியமான ஒன்றுக்கும் மேற்பட்ட இலக்கண வகைப்பாட்டைக் கொண்டிருப்பதால் அவை கோர்வையில் அவற்றின் இலக்கணச் சூழலுக்கு வேண்டி சரிபார்க்கப்படவேண்டும்; எ.கா. states என்பது படர்க்கையிட ஒருமையாக வரும் ஒரு வினையாகவோ ஒரு பன்மைப் பெயராகவோ இருக்கலாம்; அனால் முன்வரும் many என்ற பெயரடையின் சூழலில் அது பெயராகத்தான் இருக்க இயலும். பின்வருவன போன்ற பொதுவான விதிகளும் பிரெஞ்சுமொழிக்காகச் செயலாக்கபடுகின்றது: பெரும்பாலான நேர்வுகளில் ஒரு பிரெஞ்சு வினை உடனடியாக ஒரு பெயரால் தொடரப்பட இயலாது; இடையில் சார்படையோ உடைமை மாற்றுபெயரோ வரவேண்டும். எந்த எதிர்பார்ப்பும் (எ.கா. *prendre note*) ஒப்புருசொல் அகராதியில் குறிப்பிடப்பட்டுள்ளன. மயக்கம் தீர்க்கப்படாத ஒப்புருசொற்களின் நேர்வில் அடிக்கடி நேரும் இலக்கண வகைப்பாடு கருதப்படுகின்றது. இந்த 'அண்மை சூழலால் பொருள் மயக்கநீக்குவது' (disambiguation by near context) மிகவும் அண்மைக்கால இயந்திர மொழிபெயர்ப்பு மற்றும் இயற்கை மொழி ஆய்வு ஒழுங்குமுறைகளில் பொதுவாகக் காணப்படும் பகுப்பய்விலிருந்து மிகவும் வேறுபட்டது.

7. நிறுத்தற்குறிகள், இணைபுக்கிளவிகள் (எ.கா. because), சார்பு மாற்றுபெயர்கள் (எ.கா. that) போன்றவற்றைத் தேடி அறிந்து வாக்கியங்கள் முக்கிய (main) மற்றும் துணை எச்சத்தொடர்களாகப் (subordinate clauses) கூறுபடுத்தப்பட்டுள்ளன. எச்சத்தொடர்களின் (clauses) தொடக்கங்களுக்கும் இறுதிகளுக்கும் குறிகள் செருக்கப்பட்டுள்ளன.

8. பெயர்களுக்கும் விசேடணைகளுக்கும் இடையில் உள்ள உறவுகள் (articles and adjectives, possessive pronouns), பெயரடைகளுக்கும் வினையடைகளுக்கும் இடயிலுள்ள உறவுகள், வினைகளுக்கும் செயப்படுபொருள்களுக்கும் இடையிலுள்ள உறவுகள், முன்னுருபகளுக்கும் அவற்றின் 'செயப்படுபொருள்' பெயருக்கும் (பெயர்த்தொடர்களுக்கும்) இடையிலுள்ள உறவுகள் மற்றும் 'செய்ய'-வினையெச்சங்கள் (infinitives), வினைப்பெயர் (gerund) என்பனவற்றிற்கும் பிற சொற்களுக்கும் இடையிலுள்ள உறவுகள் போன்ற 'முதன்மைத் தொடரியல் உறவுகள்' (primary syntactic relations) தீர்மானிக்கப்படும்.

9. தொடர்களுக்குள் உள்ள சமநிலை அமைப்புகள் போன்ற 'கணக்கெடுப்புகளால்' உறவுள்ள சொற்கள் கண்டுபிடிக்கப்படும். எடுத்துக்காட்டாக பின்வரும் வாக்கியம் அ-வில் smogக்கும்

pollution control என்ற தொடருக்கும் சமநிலை இணைப்பை (coordination) அடையாளங் காணப்படுவேண்டும்; ஆ-வில் உள்ள smog மற்றும் pollution என்பன control என்பதன் அடைகளாக அடையாளங் காணப்படுவேண்டிய சமநிலை இணைப்பு.

1.அ. Smog and pollution control are important factors.

ஆ. Smog and pollution control is under consideration

இந்நேர்வில் தொடரியல் தகவல் உள்ளது (are மற்றும் is) ஆனால் பிற நேர்வுகளில் பொருண்மையியல்சார் தகவல் தேவை: எடுத்துக்காட்டாக வாக்கியம் 2-இல் zinc மற்றும் alluminium (கூறுகளின் வகைகளாக) என்பதன் சமநிலை இணைப்பு 'வேதியல் தனிமம்' ('chemical element') பொருண்மையியல் குறியீட்டின் பங்கீட்டால் உரிமம்தரப்பட்டுள்ளது

2. zinc and alluminium components

துணை வழக்கம் (sub routine) தொடரியல் உறவுகளைத் தீர்மானிக்கின்றது. எ.கா. வாக்கியம் 3-இல் speed என்பதற்கு முந்தைய கட்டத்தில் முன்னரே நிறுவப்பட்டுள்ள demand என்பதன் 'செயப்படுபொருள்' தகுதிநிலை accuracy என்பதற்கும் தரப்படுள்ளது.

3. The task demands speed and accuracy.

10. 'புற' தொடரமைப்புகளைத் தாண்டி 'ஆழமான' ஆய்வுக்குச் சென்று எழுவாயும் பயனிலையும் அடையாளம்காணப்படுகின்றன. இது ஒப்பீட்டளவில் எளிமையான செயற்பாங்காகும்: முன்னரே அடையாளம் காணப்பட்ட முற்று வினைகள், சாத்தியமான பயனிலைகளும் பெயர்களும் (அல்லது மாற்றுப் பெயர்களும்) ஆகும்; முன்னரே அடையாளம் காணப்பட்ட 'செயப்படுபொருள்கள்' சாத்தியமான 'எழுவாய்கள்' அல்ல. இவ்வழியில் வாக்கியங்கள் கூற்றுவாக்கியம், வினாவாக்கியம் அல்லது ஏவல் வாக்கியம் என குறிக்கப்படுகின்றன.

11. ஆழமான (வேற்றுமை (case)) உறவுகள் அடையாளம்காணப்படுகின்றன. அதாவது பயனிலைகளுக்கும் செயப்பாட்டு வாக்கியங்களில் வினையின் தருக்கச் செயப்படுபொருளாக அடையாளங்காணப்படும் இலக்கண எழுவாய்களையும் உள்ளடக்கிய பங்கேற்பாளர்களுக்கும் இடையிலுள்ள உறவுகள். சிஸ்ட்ரானின் பெரும்பாலான பதிப்புகளில் முன் உருபுகளுக்கும் (prepositions) அவற்றின் ஆளுநர்களுக்கும் (governors) (எ.கா. பெயர் அல்லது வினை) இடையில் உறவுகளை நிறுவுவதற்கு வழக்கங்களை இந்த நிலை உட்படுத்துக்கின்றது.

10 மற்றும் 11 கட்டங்கள் (நிலைகள்) இடைக்கிடையே நேரும் மற்றும் விதிவிலக்கான அமைப்புகளைக் கையாளுவதற்கு பகுப்பாய்வு அகராதிகளின் (analaytic dictionaries) கலந்தறிதலை உட்படுத்தலாம்.

மாற்றல் மூன்று பகுதிகளைக் கொண்டது.

12. 'கட்டுப்பாட்டு மரபுத்தொடர்களின்' சொல்சார் மாற்றல் (lexical transfer of 'conditional idioms'). நிலைபேறுபெற்ற மரபுச்சொற்கள் (standard idioms) மற்றும் நிலையான தொடர்கள் (fixed phrases) மரபுச்சொல் அகராதி (idiom dictionary) மற்றும் எல்லைக்குட்பட்ட பொருண்மையியல்சார் அகராதி (Limited Semantics dictionary) (கட்டம் 2 மற்றும் 5) வழி பகுப்பாய்வின் போது முன்னரே கையாளப்பட்டுள்ளன. இக்கட்டத்தில் பிற சொற்கள் நிபந்தனை பொருண்மையியல்சார் அகராதியில் (Conditional Semantic dictionary) வரையறை விளக்கம் செய்யப்பட்ட சில நிபந்தனைகளால் 'மரபுச்சொல்சார்' மொழிபெயர்ப்புகளைப் ('idiomatic translations) பெறுகின்றன.

13. முந்தைய கட்டத்தில் கையாளப்படாத முன்னுருபுகளின் மொழிபெயர்ப்பு. வினை வடிவுகளுடன் இணைக்கப்பட்டுள்ள தகவலாலோ அல்லது ஆளுமை செய்கிற (governing) அல்லது சார்பான சொற்களுடன் (dependent words) இணைக்கப்பட்டுள்ள குறியத்தாலோ தேர்வு தீர்மானிக்கப்படுகின்றது.

14. 'சொல்சார் வழக்கங்களைப்' (lexical routines) பயன்படுத்தி அமைப்பு மாற்றல். அதாவது சோதனைகள் குறிப்பிட்ட சொற்களுக்கோ சொற்களின் குறிப்பிட்ட தொடரியல் அல்லது பொருண்மையியல்சார் வகைப்பாடுகளுக்கோ அகராதியில் குறிப்பிடப்பட்டுள்ளன. எடுத்துக்காட்டாக, பிரஞ்சு மொழியில் *as* (*comme, pendent que, à mesure que, puisque*) என்பதன் பொருத்தமான மொழிபெயர்ப்புகளின் தெரிவு கட்டுமானங்கள் மற்றும் வினைக் காலங்களின் (verb tenses) மாற்றுத் தேர்வுகளைத் (altenative selections) தூண்டுகின்றது. மற்றொரு எடுத்துக்காட்டாக except என்பதற்கு ஒதுக்கப்பட்டுள்ள வழக்கங்கள் தற்சுட்டு (reflextive) மற்றும் நிபந்தனை (subjunctive) அர்த்தங்களுடன் பிரஞ்சு மொழியில் சரியான கட்டுமானங்களை உறுதி செய்கின்றது.

எ.கா.

He excepts to come

Il s'attend à ce qu'il vienne.

இறுதிக் கட்டத்தில் உருவாக்கமும் மூன்று அடிப்படைப் பகுதிகளக் கொண்டது.

15. மாற்றலின் (transfer) போது முன்னரே மொழிபெயர்க்கப்படாத எந்த சொல்லுக்கும் இயல்புநிலை மொழிபெயர்ப்பை (default translation) ஒதுக்குதல். எடுத்துகாட்டாக Station என்பதற்கு இயல்புநிலை பிரஞ்சு மொழிபெயர்ப்பு poste என்பதைக் கொண்டிருக்கும்; gare என்ற மாற்றுகள் சூழல் தகவல்களிலிருந்து முன்னரே தேர்ந்தெடுக்கப்படாவிட்டால் இது தேர்ந்தெடுக்கப்படும்.

16. முந்தைய நிலைகளிலிருந்து வேற்றுமை, பால், எண், காலம் போன்றவைப் பற்றிய அமைப்புத் தகவல் அடிப்படையிலும் திரிபுகள் மற்றும் சார்பு கட்டுப்பாடுகள் பற்றிய தகவல் அடிப்படையிலும் உருபன்சார் உருவாக்கம் (morphological generation). எ.கா. இயக்கங்களின் (motion) பிரஞ்சு வினைகள் (partir) *avoir* என்பதற்குப் பதிலாக இறந்தகாலத்தில் *être* என்ற துணைவினையுடன் வினைதிரிபாக்கம் செய்யும் மற்றும் சார்பான வினையெச்சம் (dependent infinitive) de (proposer de sortir) என்பதற்குப் பதிலாக நேரடியாகச் சில வினைகளுக்குப் (aimer allet) பின்னர் வரும்.

17. இலக்குமொழி சொல் வரிசையின் உருவாக்கம். எ.கா. ஆங்கில பெயரடை-பெயர் கோர்வையிலிருந்து பிரஞ்சு பெயர்-பெயரடை கோர்வையை சொல்வரிசையாக மறுவரிசைப் படுத்தல். பிரஞ்சில் உருவாக்கத்தில் இக்கட்டமும் அசை கெடுதலைக் கையாளும். எ.கா. le homme → l'homme மற்றும் முந்தைய வாக்கியத்தில் முற்சுட்டுப் பெயரின் இலக்கணப் பால் (grammatical gender) (ஆண்பால், பெண்பால்) குறிப்புரையால் அதன் மொழிபெயர்ப்பைக் கையாளும்.

3.6.3. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் பண்புகள்

சிஸ்ட்ரான் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை நேரடி அணுகுமுறை (direct approach) அடிப்படையிலானது என்று பண்பாக்கம் செய்யப்பட்டுள்ளது. இருப்பினும் மொழிபெயர்ப்பின் நிலைகளின் விளக்கத்திலிருந்து இவ்வொழுங்குமுறையை மாற்றல் ஒழுங்குமுறை (transfer system) என்று வகைபடுத்த இயலும். இது தனியாகப் பிரிக்கப்பட்ட ஆய்வுநிலை (analysis), மாற்றல் (transfer) மற்றும் உருவாக்கம் (synthesis) ஆகிய கட்டங்களை/நிலைகளை கொண்டிருப்பதாகத் தோன்றுகின்றது. முதல் பார்வையில் பகுப்பாய்வு

நிலைகள்/கட்டங்கள் இலக்கு மொழித் தரவின் குறிப்பிட்ட விளக்கத்தை ஏற்பதில்லை மற்றும் குறைந்தது உருவாக்கத்தின் ஒரு பாகம் மூல மொழியிலிருந்து சுதந்திரமானது. இந்தப் பண்புக்கூறுகள் தான் இதை உருவாக்கினவர்கள் முன்னால் விளக்கியதை உரிமைகொண்டாட வழிகாட்டியது.

சிஸ்ட்ரானின் அடிப்படை வடிவமைப்பு 1970களில் அறிமுகப்படுத்தப்பட்ட ரஷ்ய-ஆங்கில ஒழுங்குமுறையிலிருந்து கணிசமான அளவு மாறியது என்பதை ஒப்புக்கொள்ளவேண்டும்.

சியிசிக்கு வேண்டி ஆங்கிலம்-பிரஞ்சு ஒழுங்குமுறையின் உருவாக்கத்தால் சிஸ்ட்ரானை உருவாக்கியவர்கள் கூறுபடுத்தலின் (modularity) அதிக அளபுக்கு மாறினர்; இதன் காரணமாக ஒரு மூலமொழிக்கோ அல்லது இலக்கு மொழிக்கோ வடிவமைக்கப்பட்ட நிரல்கள் வேறு இலக்குமொழியின் அல்லது மூலமொழியின் மற்றொரு பதிப்புக்கு பரிமாற்றப்பட முடிந்தது. இதன் விளைவாக மூலமொழியான ஆங்கிலத்தின் பகுப்பாய்வின் நிரல்கள் எல்லா சிஸ்ட்ரான் ஒழுங்குமுறைகளுக்கும் பொதுவானதாய் மாறியது.

இருப்பினும் பல காரணங்களால் சிஸ்ட்ரான் ஒழுங்குமுறையை மாற்றல் ஒழுங்குமுறை என்று பண்பாக்கம் செய்ய இயலாது (Hutchins and Somers, 1992: 184).

முதலாவதாக, பகுப்பாய்வுக்கும் உருவாக்கலுக்கும் மொழியியல் தரவை ஒருமொழியத் தரவுத்தளங்களாகப் (monolingual databases) பிரிப்பதில்லை; இருமொழியத் தரவுத்தரங்கள் (bilingual databases) நேரடியான சொல்சார் மற்றும் அமைப்பு மாற்றலின் எல்லைக்கு உட்படுத்தப்பட்டது. இரண்டாவதாக, புலக்குறிப்புகள் இருந்த போதிலும் மாற்றலுக்கும் உருவாக்கலுக்கும் தெளிவான பிரிப்பு இல்லை. மூன்றாவதாக, வாகியங்களின் முழுமையான பகுப்பாய்வு இல்லை. நான்காவதாக, பகுப்பாய்வுகள் சில நேர்வுகளில் தடுக்கப்பட்டதாகத் தோன்றுகின்றது. ஐந்தாவதாக, 'அகராதிசார்' தீர்வுகளுக்கு (lexicographic solutions) எச்ச முன்னுரிமையின் (residual preference) தெளிவன சான்று இருந்தது. இறுதியாக, பொதுவான மாதிரி (modle) அல்லது சட்டக அமைப்பு (framework) குறைவின் காரணமாகப் பொருண்மையியல்சார் பண்புக்கூறுகளின் (semantic features) பிரயோகத்தில் முரண்பாட்டின் சான்று காணப்பட்டது.

சமச்சீர்மை (homogeneity) மற்றும் இணக்கத்தின் (conformity) அதிக அளபை அறிமுகப்படுத்த வெற்றிகரமான முயற்சிகள் மேற்கொள்ளப்பட்டாலும் [குறிப்பாக ரொமான்ஸ் மொழிகளுக்கு பொதுவான உடற்பகுதியை (common trunk) உருவாக்குதல்] சிஸ்ட்ரான் முதல் தலைமுறை நேரடி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் பன்முகப் பண்பை பிரதிலித்தது என்பது பொதுவான முடிவு ஆகும்.

3.7. லோகோஸ் (LOGOS)

பெர்னார்ட் ஸ்காட்டால் நிறுவப்பட்ட லோகோஸ் கார்ப்பரேஷனால் (Logos Corporation) லோகோஸ் இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பு 1964-இல் தொடங்கப்பட்டது. 1971-இல் அதன் முதல் நிறைவேற்றம் இராணுவத் தளவாடங்களின் ஆங்கிலப் பராமரிப்புக் கையேட்டை வியட்நாம் மொயில் மொழிபெயர்ப்பு செய்ய அமெரிக்க விமானப்படையால் பயன்படுத்தப்பட்டது. போரில் அமெரிக்காவின் ஈடுபாடு முடிவுற்றபோது அதன் பயன்பாடும் முடிவுக்கு வந்தது. லோகோசின் மொழியியல் அடித்தளம் வெளிப்படுத்தப்படவில்லை. இயந்திர ஒழுங்குமுறையை உருவாக்கியவர் அதன் மொழியியல் அணுகுமுறை வேற்றுமை இலக்கணம்/இணைதிறக் கோட்பாடுக்கு (case grammar/valency theory) ஒப்பான வழிகள் கண்டுபிடிக்கப்பட்டு இயற்கை மொழி கிளைப்படமாக ஒழுங்குபடுத்தப்பட்ட பொருண்மையியல்-தொடரியல் சாராம்ச மொழியுடன் பொருத்தப்பட்டது என்று கூறுகின்றார்.

லோகோஸ் தொடர்ந்து வாடிக்கையாளர்களைக் கவர்ந்தது. 1978-இல் சீமென்ஸ் எஜி (Siemens AG) தொலைத்தொடர்பு கையேடுகளை (telecommunications manuals) மொழிபெயர்க்க லாகோஸ் (LOGOS) ஜெர்மன்-ஆங்கில மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கத்திற்கு நிதியுதவி செய்யத் தொடங்கியது. மூன்றாண்டுகளுக்குப் பிறகு லோகோஸ் ஒரு உற்பத்தி ஒழுங்குமுறையைக் கொடுத்தது. ஆனால் அது பயன்பாட்டிற்கு பொருத்தமானதாக இல்லை. இறுதியாக லோகோஸ், வாங் கணினி நிறுவனத்துடன் (Wang computer company) ஒரு உடன்படிக்கை ஏற்படுத்திக்கொண்டது. அது வாங் அலுவலக கணினிகளில் ஜெர்மன்-ஆங்கில மொழிபெயர்ப்பு ஒழுங்குமுறையின் செயல்படுத்தத்தை அனுமதித்தது. இந்த இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை வணிகச் சந்தையைச் சென்றடைந்து; பல பன்னாட்டு

நிறுவனங்கள் அதை வாங்கிப் பயன்படுத்தின. ஆங்கிலம்-பிரஞ்சு, ஆங்கிலம்-ஜெர்மன் போன்ற பிற மொழி இணைகளின் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் உருவாக்கமும் தொடங்கப்பட்டன.

3.8. சூசி (SUSY: SAARBRÜCKER ÜBERSETZUNGS SYSTEM)

சூசியைப் பற்றிய ஒரு சுருக்கமான விளக்கம் இங்கு தரப்பட்டுள்ளது (Hutchins and Somers, 1992: 192-205)

3.8.1. பின்னணி

ஜெர்மனியில் சார்ப்ரூகனில் (Saarbrücken) உள்ள சார் பல்கலைக்கழகம் பிந்தைய 1960களில் நிறுவப்பட்ட ஐரோப்பாவில் மிகப்பெரிய இயந்திர மொழிபெயர்ப்பு ஆய்வுத்திட்டங்களில் ஒன்றிற்கு ஆதரவளித்து (Slocum, 1985).

ரஷியன்-ஜெர்மன் மொழிபெயர்ப்புக்கு வேண்டி GAT-ஐ மாற்றும் திட்டம் தோல்வியடைந்த பின்னர் ரஷ்யன்-ஜெர்மன் மொழிபெயர்புக்கு வேண்டி ஓரளவுக்கு அதே முறையில் மாற்றல் அணுகுமுறையைப் பின்பற்றி சார்புக் கிளை அமைப்புகளாக "உலகமயமான" வாக்கியப் பகுப்பாய்விற்குப் பின்னர் ஒரு புதிய ஒழுங்குமுறை வடிவமைக்கப்பட்டது. பிற இயந்திர மொழிபெயர்ப்பு ஆய்வுத்திட்டங்களைப் போலல்லாமல் சார்ப்ரூகன் குழு பயன்பாடுகளை உற்பத்திச் செய்வதற்குக் கட்டாயப்படுத்தப்படாமல் ஆய்வு விருப்பங்களைத் தொடர்வதற்கு ஒப்பீட்டளவில் சுதந்திரமாக செயல்பட விடப்பட்டது; மட்டுமன்றி ஒரு நிலையில் நடந்துகொண்டிருந்த பரிசோனையையும் மாற்றத்தையும் அனுமதிக்க நிதியுதவி செய்யப்பட்டது. இதன் விளைவாக சூசி (SUSY) இயந்திர மொழிபெயர்ப்பு மற்றும் செயற்கை அறிவின் வெளி வளர்ச்சிகளைப் நெருக்கமாகப் பின் தொடர வாய்ப்பாக இருந்தது. எடுத்துக்காட்டாக சார்ப்ரூகன் LEIBNIZ கூட்டுறவு இயந்திர மொழிபெயர்ப்புக் குழுவை நிறுவதற்கு உதவியது. 1975 வரை சூசி (SUSY) தீவிரமான மாற்றல் அணுகுமுறையை அடிப்படையாகக் கொண்டிருந்தது; 1976இலிருந்து "ஆழமான" பகுப்பாய்வை வேண்டும் மொழியியல் சிக்கல்கள் மாற்றல் உருப்படுத்தங்களை இடைமொழியின் சில பொதுமைகளை மேற்கொள்ள வேண்டி கட்டாயப்படுத்தியதால் அது நுண்மையாக/அருவமாகப் படிப்படியாக மாறியது. மேலும் அம்மாதிரியானா ஆராய்ச்சிச்

சுதந்திரத்தின் காரணமாக குறிப்பிட்ட இறுதிப் பயனர் பயன்பாடுகள் உருவாக்கும் தொடர்ச்சியான முயற்சி இல்லாதிருந்தது.

ஆங்கிலம், ஜெர்மன், ரஷ்யன், எஸ்பிராந்தோ இவற்றை உட்படுத்திய பன்மொழிய ஒழுங்குமுறையாக உருவாக்கப்பட்ட சூசி, ரஷ்யனிலிருந்து ஜெர்மனுக்கு மொழிபெயர்க்கக் கவனம் செலுத்தியது. சூசியின் கூடுதலான வளர்ச்சிக்கு வலுவான கட்டுப்படுத்தும் காரணி அதன் ஆரம்பகால உத்வேகத்துடன் தொடர்புடையதாய்த் தோன்றுகின்றது: சூசி மொழியியல் விதிகள் தீவிரமாக சுதந்திரமான நிலைகளாக ஒழுங்குபடுத்தப்படும் பழமையான அணுகுமுறையைப் பின்பற்றியது; இதனால் அதன் திறன் மென்பொருளில் நேரடியாக உட்படுத்தப்பட்டது. இதன் விளைவாக விதிகள் கிட்டத்தட்டப் படிக்க இயலாததாய் இருந்தது மற்றும் நிர்வகிக்க இயலாததாய் மாறியது. பயன்பாட்டுச் சாத்தியம் அடிப்படையில் சூசி பயன்படுத்தப்பட்டாலும் தோல்வியுற்றது. இரண்டாம் தலைமுறை திட்டமான சூசி II 1981 தொடங்கப்பட்டது.

3.8.2. **அடிப்படை ஒழுங்குமுறை வடிவமைப்பு**

சூசி அடிப்படையில் ஒரு மாற்றல் ஒழுங்குமுறையாகும் (transfer system); இது ஒருமொழிய பகுப்பாய்வு மற்றும் உருவாக்கம் என்ற கட்டங்களையும் சொல்சார் மற்றும் அமைப்புசார் மாற்றல் நடைபெறும் இருமொழியக் கட்டத்தையும் கொண்டுள்ளது. கணினியாக்கம் செய்யப்பட்ட அமைப்புகள் அடிப்படையில் சார்பு கிளையமைப்புகள் ஆகும் (dependency tree structures); இது ஜெர்மன் மொழியியலாளர்களுக்கு இடையில் இணைதிறன் இலக்கண (valency grammar) அணுகுமுறை மேலோங்கி இருந்ததன் பிரதிபலிப்பாகும். உள்ளீட்டை விருப்பாக முன் திருத்தியமைக்க இயலும். 'தோல்வி-மென்பொருள்' (fail-soft) தொகுதிகளின் இணைப்பு காரணமாக சூசி எப்பொழுதும் ஒருவித வெளியீட்டை உருவாக்கும்.

இந்த ஒழுங்குமுறை ஒரு முக்கிய அம்சத்தில் அதிகத் தொகுதித்தன்மை (modular) உடையதாகும்: மொழிபெயர்ப்புச் செயற்பாங்கு தனித்தனியான துணைச் செயற்பாங்குகளாகப் பிரிக்கப்பட்டுள்ளன. தொகுதிகள் (modules) கண்டிப்பான வரிசைமுறையில் பயன்படுத்தப்படுகின்றது. பொதுவாக, பகுப்பாய்வு மற்றும் உருவாக்கத் தொகுதிகள் (modules) மொழித் திட்டவட்டமானவை; அதே வேளையில் மாற்றல் தொகுதிகள் தனிப்பட்ட மொழி இணைகளுக்காக வடிவமைக்கப்பட்டவை. இருப்பினும் நாம் விளக்கும் பதிவில் சில தொகுதிகள், குறிப்பாகப் பகுப்பாய்வில், குறிப்பிட்ட மொழி இணைகளிலிருந்து சுதந்திரம் இல்லாதனவும் உள்ளன. இந்த அளவில் சூசி ஒரு சுத்தமான பன்மொழிய மாற்றல் ஒழுங்குமுறை (multilingual transfer system) அல்ல; இதில் மூல மொழியின் பகுப்பாய்வுத் தொகுதி இந்த

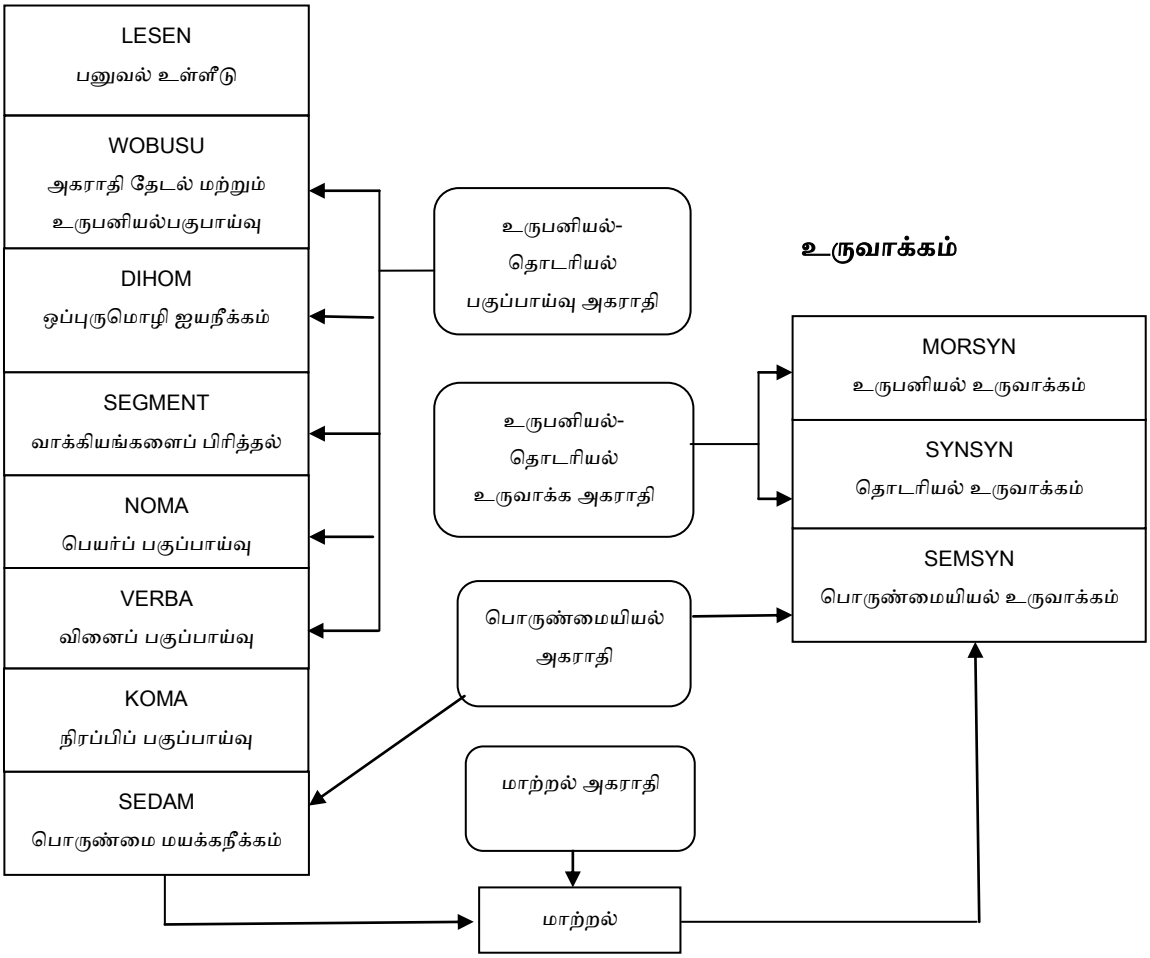
ஒழுங்குமுறையிலுள்ள எந்த இலக்கு மொழியுடனும் இணையவேண்டும். சூசி பகுப்பாய்வுத் தொகுதிகளின் நெருக்கமான பரிசோதனை அவை இலக்குமொழியையும் கணக்கில் எடுத்துக்கொள்கின்றது என்று அறிய முடிகின்றது. இது ஏன் இலக்கியத்தில் சுசியின் வேறுபட்ட பதிப்புகள் அவற்றின் விவரங்களில் சிறிது வேறுபடுகின்றன என்பதை விளக்கும். இதன் விளைவாக சூசியின் பன்மொழியத்தின் கொள்கையைக் குறைகூறவியலும் என்றாலும் சூசியின் தொகுதித்தன்மையை ஒரு நேர்மறைப் பண்புக்கூறாகப் பார்க்க இயலும். இது பகுப்பாய்வு மற்றும் உருவாக்கச் செயற்பாங்கின் சில பகுதிகளில் திருத்தங்களையும் மாற்றகளையும் அனுமதிக்கின்றது; அதேசமயம் முந்தைய பதிப்பின் பிற பகுதிகளைத் தக்கவைத்துக்கொள்ளும்; இதன் காரணமாகப் புதிய மொழியியல் கருத்துக்களையும் அணுகுமுறைகளையும் இணைத்துக்கொள்ள இயலும்; ஒரு மொழி இணையின் பகுப்பாய்வுத் தொகுதிகளை வேறுபட்ட மொழி இணைகளுக்குச் சுவீகரித்துப் பயன்படுத்தவும் இயலும்.

வேறு அம்சத்தில் சூசி அதிக மதிப்பெண் பெறவில்லை. பதிப்புகளுக்கிடையில் வேறுபாடுகள் இருந்தாலும் பொதுவாக மொழியியல் மற்றும் வழிமுறை நோக்குகளின் பிரிவினை தெளிவாக இல்லை. அதிகமாகத் தொகுதியாக்கம் செய்யப்பட்டிருந்தலும் ஆரம்ப கட்ட சூசி ஒழுங்குமுறை பிரிவினையைத் திறம்படக் கொண்டிருக்கவில்லை; பல்வேறு தொகுதிகளும் நேரடியாக ஃபோர்ட்ரானில் நிரல்செய்யப்பட்டுள்ளன.

சூசியில் செயலாக்கத்தின் அடிப்படை நிலைகள் கீழே தந்துள்ள படத்தில் காட்டப்பட்டுள்ளன (Hutchins and Somer, 1992: 193):

சூசியின் மொழிபெயர்ப்புச் செயற்பாங்கு

பகுப்பாய்வு



மாற்றல்

தலைப்புகளுடனான துணைநிலைகள் (சூசியின் கலைசொலின் படி 'இயக்கிகள்' (operators)) பின்வருமாறு அமையும்: பனுவல் உள்ளீடு (LESEN) முன் திருத்தியமைத்தலின் விருப்பநிலையால் முன் தொடரப்படலாம். அகராதி தேடலும் (WOBUSU) அதனுடன் தொடர்புடைய உருபனியல் பகுப்பாய்வும் ஒப்புருமொழி ஐயநீக்கத்தால் (DIHOM) பின்தொடரப்படும்; அதன் பின்னர் தொடரியல் பகுப்பாய்வின் வேறுபட்ட நிலைகளால் பின் தொடரப்படும்: SEGMENT எச்சத்தொடர்களையும் தொடர்களையும் மற்றும் தற்காலிக சார்பு அமைப்பையும் அடையாளங்காணும். NOMA பெயர்க் குழுக்களையும் VERBA வினைக் குழுக்களையும்

அடையாளங்காணும். முழு வாக்கியமும் KOMAவால் முழுமையான பகுப்பாய்வைப் (parse) பெறும்; சொல் மயக்கமும் அமைப்பு மயக்கமும் SEDAM என்ற பொருண்மை மயக்கநீக்கத் தொகுதியால் கையாளப்படும். தனிநிலையான மாற்றல் (TRANSFER) சொல் மற்றும் அமைப்பு மாற்றலை உள்ளடக்கும். பொருண்மையியல் உருவாக்கம் (SEMSYN), தொடரியல் உருவாக்கம் (SYNSYN), உருபனியல் உருவாக்கம் (MORSYN) என்ற நிலைகளைக் கடந்து இலக்கு மொழி உருவாக்கப்படும்.

மூல மொழிக்கும் இலக்கு மொழிக்கும் தனித்தனியான அகராதிகள் உள்ளன; அவை சொற்களுக்கு இன்றியமையானதாக உருபனியல் மற்றும் தொடரியல் தரவுகளைக் கொண்டுள்ளன. ஒவ்வொரு மொழியும் பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் தனித்தனியான அகராதிகளைக் கொண்டுள்ளன. மிகப்பெரியவை140,000 பதிவுகளைக் கொண்ட ஜெர்மன் பகுப்பாய்வு அகராதியும் 14,000 பதிவுகளைக் கொண்ட ஜெர்மன் உருவாக்க அகராதியும் ஆகும். பொருண்மையியல் செயற்பாங்குகள் மூல மொழிக்கும் இலக்கு மொழிக்கும் பொருண்மையியல் அகராதிகளை ஈடுபடுத்தும். இந்நேர்வில் அதே ஒருமொழிய அகராதி பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் பயன்படுத்தப்படும்; இதில் மிகப்பெரியது 75,000 பதிவுகள் உள்ள ஜெர்மன் அகராதியாகும். பொருண்மையியல் அகராதிகள் பொருண்மைப் பண்புக்கூறுகளையும் தொடரியல் மற்றும் பொருண்மையியல் கட்டுப்பாடுகளையும் கொண்டிருக்கும்; குறிப்பாகப் பின்னுருபுகளுக்குக் கொண்டிருக்கும். இருமொழிய மாற்றல் அகராதிகள் அடிப்படையில் அமைப்புசார் சூழலைப் பற்றிய சில தகவலுடன் சொல் நிகரன்களையும் கொண்டிருக்கும். அவற்றிம் ஒப்பீட்டளவிலான சிறிய அளவுகள் (ஆங்கிலம்-ஜெர்மன் 10,000க்கும் மேற்பட்ட பதிவுகள்) சூசி ஆய்வுத்திட்டத்தின் பரிசோதனை இயல்பை பிரதிபலிக்கும்.

இங்கு சூசியில் காணப்படும் தரவு அமைப்பின் வகை, முன் திருத்தியமைப்பு செயன்மைகள், ஏதாவது உட்கூறு தோற்றால் ஏதாவது சில வெளியீட்டை உறுதிசெய்யும் இயக்கமுறை (மீட்பு இயக்கி (RESCUE OPERATOR)) இவற்றைப் பற்றிய சுருக்கமான விளக்கத்தைத் தந்துவிட்டு தொகுதிகளின் செயன்மைகள் விளக்கப்படும்.

3.8.3. **தரவு அமைப்பு**

எந்த தொகுதிமயமான ஒழுங்குமுறையில் (modular system) உள்ளது போன்று தரவு அமைப்பு முக்கியமானதாகும்; ஏனென்றால் இது தொகுதிகளுக்கிடையே பரிமாற்றத்தை

உறுதிசெய்கின்றது. இதை நாம் இரு கண்ணோட்டத்தில் கருதவியலும்: கணினி சார்ந்த மற்றும் மொழியியல் சார்ந்த கண்ணோட்டங்கள். கணினிசார் கண்ணோட்டத்தின் படி, சூசியின் தரவு அமைப்பு சிஸ்ட்ரானில் காணப்படுவதை ஒக்கும். இது நிரலாக்க வசதியால் ஊக்கபடுத்தப்பட்டுள்ளது; சூசியில் நிரலாக்க மொழி ஃபோர்ட்ரான் ஆகும். ஒரு சொல்லுக்கு ஒரு தரவுப் பதிவு (data record) இருக்கும் போது சூசி தரவு அமைப்பைப் பயன்படுத்தும்; அவை ஒரு வரிசைமுறையிலான கோப்பில் சேமிக்கப்படும். 'சொல் பதிவு' ஒன்றோ அல்லது அதற்கு மேலோ பைடில்/இருமஎண்ணில் (bite) ஆன 'செல்கள்'ஆகப் பிரிக்கப்பட்டுள்ளது. மொழிபெயர்ப்புச் செயற்பாங்கு பொருத்தமான செல்லில் மதிப்புகளை மாற்றி தொடர்ச்சியாகக் கோப்பை விரித்தெழுதுவதைக் கொண்டிருக்கும்.

மொழியியல் கண்ணோட்டத்தில் தரவு அமைப்பு ஒரு சார்பு கிளை அமைப்பு (dependency structure) ஆகும்; இதில் ஒவ்வொரு சொல்லுக்கும் ஆள்வான்-சார்வான் (governor-dependent) உறவுகள் அடையாளம் காணப்படும். the very big system என்ற பெயர்த்தொடரில் system ஆள்வான் ஆகும்; அது the மற்றும் big என்ற அண்மைச் சார்வான்களை கொண்டுள்ளது; big என்ற பெயரடை very என்ற சார்வானைக் கொண்டுள்ளது. வாக்கிய நிலையில் முக்கிய வினை ஒட்டுமொத்தமான ஆள்வானாக எடுத்துக்கொள்ளப்படுகின்றது; அது (எழுவாய், பயனிலை என்ற) நிரப்பிகளையும் அடைகளையும் (சூழ்நிலை வினையடைகள்) சார்பான்களாகக் கொண்டிருக்கும். சார்புத் தகவல் சார்புகளின் வகைகளின் அடையாளங்களால் பெரிதுபடுத்தப்பட்டிருக்கும், எ.கா. எழுவாய் அல்லது செயப்படுபொருள், அடை (modifier) அல்லது அடைகொளி அடை (determiner), போன்றன.

சிஸ்ட்ரானும் மற்றும் சூசியும் முரட்டுத்தன்மையான கீழ் நிலை தரவு அமைப்புகளைப் பயன்படுத்தினாலும் மொழியியல் விளக்கம் சூசியில் அதிநவீனமானதாகும். குறிப்பாக, செல்களில்/அறைகளில் சில 'சுட்டிக்காட்டிகள்' (pointers) ஆகப் பயன்படுத்தப்பட்டுள்ளன; இது பின்வரும் வழியில் கிளையமைப்புகளை உருப்படுத்தம் செய்வதற்குப் பயன்படுத்த வரிசைமுறையிலான கோப்பு அமைப்பை அனுதிக்கும். ஒவ்வொரு சொல் பதிவுக்கும் வாக்கியதில் அது உருப்படுத்தம் செய்யும் சொல்லின் இருப்பிடம் அடிப்படையில் தொடர்ச்சியாக எண் தரப்பட்டுள்ளது. சொல் பதிவில் ஒவ்வொரு இருமஎண்ணும் ஒவ்வொரு தனியான எழுத்துருவை சேமிக்க இயலும்: ஒரு எழுத்தோ முழு எண்ணோ. முதல் செல்/அறை 20 இருமஎண்களைக்

கொண்டிருக்கலாம் மற்றும் அகராதியிலிருந்து இலக்கணத் தகவல்களைக் கொண்டிருக்கலாம் [எ.கா. வகைப்பாடு N = noun (பெயர்), V = verb (வினை) போன்றன மற்றும் பெயர் என்றால் S(ingular) (ஒருமை) அல்லது P(lural) (பன்மை) அல்லது வினை என்றால் அதன் மூவிடம், எண் மற்றும் காலம் போன்றன]

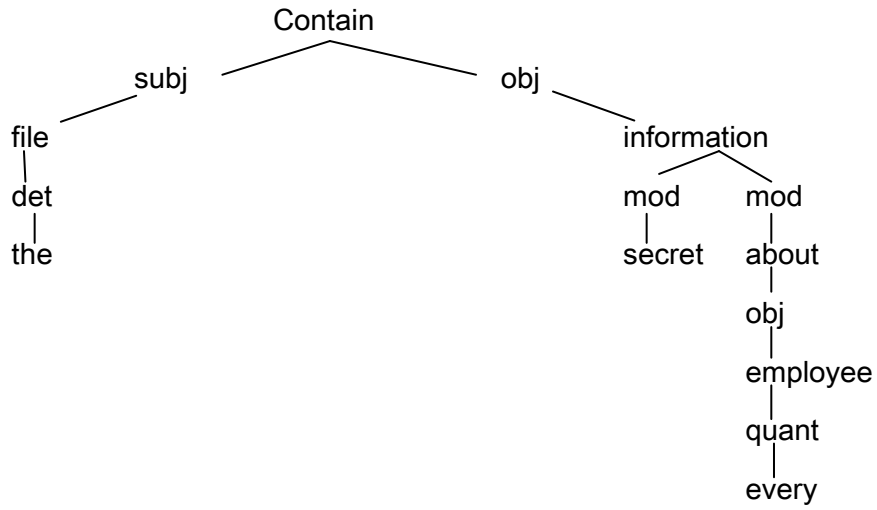
எடுத்துக்காட்டிற்காக 75 மற்றும் 76 என எண்ணிட்ட இருமஎண்கள் (bites) சுட்டுக்காட்டித் தகவலைச் சேகரிக்கப் பயன்படுத்தப்படுகின்றது என்று அனுமானிப்போம்: இருமஎண் 75இல் தற்போதைய பதிவின் ஆளுவானின் (governor) பதிவு எண் தோன்றும்; மற்றும் இருமஎண் 76இல் அது எப்படிப்பட்ட சார்பான் (dependent) என்பதைக் காட்டும் குறிமம் [எ.கா. S(ubject) (எழுவாய்), O(bject (செயப்படுபொருள்), M(odifier) (அடை) போன்றன]. இந்த வழியில் வாக்கியம் 1 க்கு பின்வரும் படத்தில் காட்டப்பட்டுள்ள தரவு அமைப்பு நமக்கு இருக்க இயலும் (Hutchins and Somer, 1992: 195):

(1) The file contains secret information about every employee.

படம் தரவு அமைப்பு (உருவகப்படுத்தப்பட்டது)

Record no 1 2 3 4 5 6 7 8 9 10 11	12	13	14	15	16	17	18	19	20	21	22	23	23	34	35	36	37	38	39	40
1	THE									D									THE	2D
2	FILE									NS									FILE	3S
3	CONTAINS									V3SF									CONTAINS	0
4	SECRET									A									SECRET	5M
5	TRANSFORMATION									NS									INFORM	3O
6	ABOUT									p									ABOUT	5M
7	EVERY									Q									EVERY	8Q
8	EMPLOYEE									NS									EMPLOY	6O

இருமஎண்கள் 75 மற்றும் 76இல் உள்ள தகவல்கள் பின்வரும் சார்புக் கிளையமைப்பை திறம்பட உருப்படுத்தம் செய்கின்றது:



தரவு அமைப்பு முரட்டுத்தன்மையானது என்றாலும் இந்தச் சுட்டிக்காட்டும் இயங்குமுறையைப் பயன்படுத்தி இதை மொழியியல் அடிப்படையில் அதிநவீன முறையில் பயன்படுத்த இயலும். இருப்பினும் மாற்று பகுப்பாய்வு சாத்தியமாகும் போது (எ.கா. நாம் எடுத்துக்காட்டும் வகைப்பாட்டுப் பொருண்மை மயக்கங்கள் (category ambiguities)) சிக்கல் ஏற்படும். மாற்று கிளையமைப்பை உருவாக்குவது ஒன்றோ இரண்டோ குறியத்தின் விவரங்களில் மட்டும் வேறுபடும் சொல் பதிவு கோப்புகளின் நகல்கள் இருக்கவேண்டும் என்று அர்த்தமாகும்.

3.8.4. **முன் திருத்தியமைத்தல் மற்றும் தோல்வி-மென்பொருள் இயக்கி**

சூசி பகுப்பாய்வுக்கு உதவ சிறப்பான குறியங்களைச் செருகி மூலப் பனுவல்களை முன் திருத்தியமைப்பதின் (pre-editing) விருப்பத்தேர்வை தருகின்றது. முன் திருத்தியமைப்பு ஒழுங்குமுறையின் செயல்திறன் அதிகரிப்பதை அனுமதிக்கின்றது; ஆனால் இது அதன் தேவையான பண்புக்கூறு அல்ல. பகுப்பாய்வு செயற்பாங்குகள் இந்தக் குறியங்களுக்கு வேண்டி எட்டிப்பார்க்கும்; ஆனால் அவை இல்லாமல் இயங்க இயலும்.

சூசியின் தனித்தன்மை மற்றும் புதுமையான பண்புக்கூறு தோல்வி-மென்பொருள் (fail-soft) இயக்கியான மீட்பு (RESCUE) என்பதன் இணைப்பாகும்; ஏதாவது தொகுதியில் ஒன்று

தவறாகப் போய்விடும் போதெல்லாம் மீட்பு இயக்கத்திற்கு வரும். ஒவ்வொரு தொகுதியும் உள்ளீடும் வெளியீடும் எதிர்பாப்புக்கு இணக்கமாக இருக்கின்றதா என்று பார்க்க வேண்டி நிலைத்தன்மை சரிபார்ப்புகளை உட்படுத்தும்.

இந்த அம்சத்தில் சூசி 'ஏதாவது முயல்' (try anything) ஒழுங்குமுறையாகும். உள்ளீடு எதுவாக இருந்தாலும் இந்த ஒழுங்குமுறை ஒரு வகையான மொழிபெயர்ப்பை உருவாக்கும்; அது அறைகுரையான சொல்லுக்குச் சொல் பதிவாகக்கூட இருக்கலாம்.

3.8.5. **பகுப்பாய்வு**

3.8.5.1. **பனுவல் உள்ளீடு மற்றும் அகராதி நோக்கீடு**

LESEN ('படி') என்ற முதல் தொகுதி மொழியியல் அல்லாத முன் பகுப்பாய்வு ஆகும். LESEN பனுவலைத் தொடக்கத் தரவு அமைப்பாகப் படிக்கும்; பனுவலை வாக்கியங்களாகப் பிரிக்கும்; ஒவ்வொரு சொல்லுக்கும் அடையாள எண்ணை ஒதுக்கும். இது மேலும் பனுவலைச் சீராக்கும்; இதன் அர்த்தம் அச்சுக்கலைத் தகவலை கணக்கில் எடுப்பதாகும் [குறிப்பாக ஜெர்மன் சொல்லுக்கு அது முகட்டெழுத்தில் (capital letter) தொடங்குகிறதா என்பது]. ஒவ்வொரு சொல்லுக்கும் முன்னர் கூறியது போல் 'சொல் பதிவு' உருவக்கப்படும்; இது முன் திருத்தியமைப்புத் தகவலையும் உட்படுத்தும்.

LESEN-இன் வெளியீடு WOBUSU (Wörterbuchsuche 'dictionary look-up') என்ற அடுத்தத் தொகுதிக்கு அனுப்பப்படும். முன் திருத்தியமைப்பின் போது இயற்பெயராக அடையாளப்படுத்தப்பட்டவை தான் அலட்சியப்படுத்தப்படும். சூசியில் மூன்று ஒரு மொழிய அகராதிகள் உள்ளன: ஒரு இலக்கணச் சொற்களின் உயர் நிகழ்வெண் அகராதி, எல்லா ஒழுங்கற்ற வடிவங்களையும் பகுதிகளையும் (stems) உள்ளடக்கிய பகுதி அகராதி (stem dictionary) மற்றும் மரபுச்சொல் அகராதி (idiomatic dictionary). அகராதிப் பதிவு, பகுதி (stem) அல்லது முழு வடிவு, சொல்லன் (lemma) (அல்லது சொல்) மற்றும் தொடர்புடைய உருபனியல் மற்றும் அடுக்குநிலைத் தகவல்களைத் (paradigmatic information) தரும். அது வகைப்பாட்டிற்கு ஏற்ப சில அடுக்கு நிலைத் தகவல்களையும் தரும். எடுத்துக்காட்டாக ஒரு பெயர்சொல்லுக்கு அதன் பால், பன்மை வடிவம், முன்னுருபு இணைவுத் திறன் (prepositional valency) (Lust AUF Kuchen ‘desire for cakes’, Vertrag MIT Guinea ‘treaty with Guinea’) போன்றவை கருத்தத்தக்கன ஆகும். வினைக்கு இந்த அகராதி அது பிரிக்கவியலும் முன்னொட்டைக் கொண்டிருக்கிறதா, அது எந்தத் திரிபு

அடுக்குநிலைகளைச் (inflectional paradigms) சாரும், எந்த வினைவகையைச் சாரும் (எ.கா. முழு வினை, துணைவினை, வினைநோக்கு வினை), அது zu என்பதுடன் இணையுமா, செயப்பாட்டின் (passive) எந்த வடிவம் (ஏதுமில்லை, மூவிடம்சார். மூவிடம்சாரா) உள்ளது, இறந்தகால வினை எச்சத்தை வினையடையாகப் பயன்படுத்த இயலுமா, பொருத்தமானது என்றால் அதன் தற்சுட்டு மாற்றுப் பெயர் எந்த வேற்றுமையை ஏற்கும், கூட்டுக் காலங்களில் எந்தத் துணைவினை (haben அல்லது sein) காணப்படுகின்றது, அதன் இணைதிறன் சட்டம் எது என்பனவற்றைச் சுட்டிக்காட்டும்.

3.8.5.2. **உருபனியல் பகுப்பாய்வு**

WOBUSU தொகுதியும் உருபனியல் பகுப்பாய்வை இணைத்துக்கொள்ளும்; இது இரண்டு துணைத் தொகுதிகளாக பிரிக்கப்பட்டுள்ளது: திரிபு (INFLECTION) மற்றும் கூட்டு (COMPOUND).

முதல் தொகுதியான திரிபு தனிச் சொல்லைக் கையாளும்; மற்றும் பகுதி + ஒட்டு என்பதன் அடிப்படையில் ஒரு பகுப்பாய்வைத் தர முயலும். இது சொல்லின் எல்லாச் சாத்தியமான கூறுபடுத்தலையும் பரிசோதிக்கும்; மற்றும் பகுதிகளின் மற்றும் ஒட்டுக்களின் சாத்தியமான பிணைப்பு அல்லாதவைகளை (அதாவது முன்மொழியப்பட்ட பகுதியோ ஒட்டோ இல்லையென்றால்) வடிகட்டி வெளியேற்றும். எடுத்துக்காட்டாக speichern என்ற சொல் பின்வருமாறு பிரிக்கப்படும்:

(3) SPEICHERN + 0 SPEICHER + N + N SPEICHE + RN
SPEICH + ERN SPEIC + HERN SPEI + CHERN
SPE + ICHERN SP + EICHERN S + PEICHERN

பகுதிகள் (+ என்பதற்கு முந்தைய பாகங்கள்) அகராதியில் பார்க்கப்படுகின்றது; நான்கு மட்டும்தான் சாத்தியமானதாகக் கண்டுபிடிக்கப்படுகின்றது: அவையாவன SPEICHERN (*speichern* 'to store' என்ற வினையின் வினையெச்சம்), *SPEICHER* (Speicher 'store' என்ற பெயரின் பகுதி அல்லது Speichern என்ற வினையின் பகுதி வடிவம்) மற்றும் SPEICHE (Speiche 'spoke' of a wheel' என்ற பெயரின் பகுதி). இந்த நான்கு சாத்தியங்களில் +RN என்பது சாத்தியமான இறுதி அல்லாததால் SPEICHE+RN என்பதை உடனடியாகத் தள்ளலாம்.

(4)SPEICHERN + 0 *speichern* என்பதன் எச்ச வடிவம்

SPEICHER+ N *Speicher* என்ற பெயரின் கொடைவேற்றுமை பன்மை வடிவம்

SPEICHER+ N *speichern* என்ற வினையின் திரிபு வடிவமாகும்

இந்த பொருண்மை மயக்கத்தின் தீர்வை பிந்தைய தொகுதிக்கு ஒதுக்கப்பட்டுள்ளது. இந்த எடுத்துக்காட்டில் மூன்று மாற்றுகள் உள்ளன; ஆனால் INFLECTION என்ற பன்மடங்கு வெளியீட்டு வடிவம் ஜெர்மனில் அடிக்கடி நிகழ்வதாகும்; இது நாம் கண்டபடி சூழலுக்கு வெளியே மிகவும் பொருண்மை மயக்கமுள்ள பல இறுதிகளைக் கொண்டிருக்கும்.

திரிபுத் தொகுதி (INFLECTION) குறைந்தது ஒரு தீர்வையும் தராவிடில் கூட்டுத் தொகுதி (COMPOUND module) உயிர்ப்பிக்கப்படும். இது கூட்டுச் சொற்களையும் ஆக்கங்களையும் கையாளும். ஆங்கிலத்தில் கூட்டாக்கம், அடிப்படையில் தொடரியல் நிகழ்வாகும்; இது வரிசைமுறையில் பயன்படுத்தப்பட்டுள்ள அடையாளங் காணவியலும் சொற்களைக் கருத்தில் கொள்ளும். ஜெர்மன் மொழியில் இது உருபனியல் சிக்கலாகும். கூட்டுக்களின் தனிபட்ட தனிமங்கள் தனிச் சொற்களை உருவாக்கச் சேர்க்கப்படும்; அவை அவ்வாறே அகராதியில் அரிதாகத்தான் பட்டியலிடப்படும்.

கூட்டுத் தொகுதி (COMPOUND module) திரிபுத் (INFLECTION) தொகுதியைப் போன்றே செயல்படும்; ஆனால் கூறுகளாகப் பிரித்தல் கூடுதல் கலவைத் தன்மையானதாகும்: வெறுமனே பகுதி + ஒட்டு என்பதை கண்டுபிடிப்பதல்ல; ஒட்டுக்கள் மற்றும் வேர்களின் பன்மடங்கு அமைப்பொழுங்குகளைக் கண்டுகொளவதாகும்:

(P (Z)) R (S)

(P (G) R (S)

(P) R ((F) R)^{n} (S) n≥ 1

இதில் அடைப்புக்குறிகள் விருப்பையும் n மீண்டும் நிகழ்வதயும் குறிப்பிடும்; மற்றும்

P = முன்னொட்டு (prefix)

S = பின்னொட்டு (suffix)

R = வேர் (root)

F = கூட்டாக்க அலகு e, es, s, n

G = முன்-உருபன் ge

Z = zu என்ற உருபன்

எடுத்துக்காட்டாக (5)

(4) *herauszubringen* ('to bring out') = *HERAUS* + *ZU* + *BRINGEN* (P Z R)

ausgebildete ('educated'+ inflection) = *AUS* + *GE* + *BILDET* + *E* (P G R S)

Geburtstagsgeschenke ('birthday parents') =

GEBURT+*S*+*TAG*+*S*+*GESCHENK*+*E* (R F R F R S)

திரிபும் (INFLECTION) கூட்டும் COMPOUND ஒரு பகுபாய்வைப் பரிந்துரைசெய்யத் தவறினால் அந்த சொல் 'அறியப்படாத சொல்' என எடுத்துக் கொள்ளப்படும். இருப்பினும். இருப்பினும் ஜெர்மன் மொழி அதிக அளவில் திரிபுறுவதாகும்; பெரும்பாலும் அறியப்படாத சொற்களின் தொடரியல் வகைப்பாட்டை அவற்றின் ஈற்றலகுகளைக் கொண்டு ஊகிக்க இயலும். இவ்வாறு இறுதிகளின் மற்றும் சரிபார்க்கப்படாத பகுதிகளின் சாத்தியமான சேர்க்கைகளைத் தரும் ஏதாவது ஒரு தொகுதியின் விளைவுகள் மேலும் மட்டங்களுக்குக் கடத்த இயலும். எடுத்துக்காட்டாக *te* என்ற இறுதி, *pflegte* ('cared for') என்பதில் உள்ளது போல் வினை முற்றைக் குறிப்பிடும்; ஆனால் அவ்வாறின்றியும் இருக்கும் (எ.கா. Kette 'chain')

திரிபு (INFLECTION) மற்றும் கூட்டின் (COMPOUND) விளைவுகள் WOBUSUவின் இறுதித் துணைத் தொகுதிக்கு அனுப்பப்படும்; இது நிர்ணயிக்கப்பட்ட தொடர்களை (fixed phrases) அடையாளங்காண்பதைக் கையாளும். நிர்ணயிக்கப்பட்ட தொடர்களை அடையாளங்காணும் முன் உருபனியல் பகுப்பாய்வின் செயன்மை சூசியை பகுதி-நிர்ணயிக்கப்பட்ட மரபுச்சொற்களைக் (semi-fixed idioms) கையாள அனுமதிக்கும்; மரபுச்சொற்கள் இலக்கணத் திரிகளுக்கு ஆளாகும், எ.கா. பெயர்த்தொடர்கள் வேறுமை உருபுகளை ஏற்கும்; மரபுச்சொல்சார் வினைகள் (idiomatic verbs) காலத்திற்கும் உடன்பாட்டிற்கும் (agreement) வேண்டி திரிபுறும்.

3.8.5.3. **ஒப்புருமொழியின் பொருண்மை மயக்க நீக்கம்**

WOBUSU-இலிருந்து விளையும் தரவு அமைப்பு அகராதியில் தேடப்பட்ட சொற்களின் வரிசையாகும்; அவற்றில் பல பொருண்மைமயக்கம் உள்ளவை. அடுத்த கட்டம் DIHOM (திஹோம்) (Disambiguierung von Homographen 'homograph disambiguation' (ஒப்புருமொழி பொருண்மைமயக்கநீக்கம்)) இந்தப் பொருண்மைமயக்கங்களைத் தீர்ப்பதற்கு முயற்சிசெய்யும். பயன்படுத்தப்பட்ட உத்தி பொருந்தக்கூடிய மற்றும் பகிர்மான தகவல் (compatibility and

distributional information) அடிப்படையில் ஆகும்; எனவே இது மரபு இலக்கண அடிப்படையிலான அணுகுமுறைக் காட்டிலும் தோராயப் (ஸ்டொகாஸ்டிக்) பகுப்பாய்வு (Stochastic parsing) போன்று அமைந்துள்ளது (Hutchins and Somer, 1992: 198).

DIHOM (திஹோம்) மூன்று துணை செயற்பாங்குகளைக் கொண்டது. சாத்தியமற்ற வரிசைமுறைகளையும் சாத்தியமானவற்றின் தரவரிசைகளையும் கையாளும். இரண்டு பொதுவான வழக்கமுறைகள் (routines), தனித்தன்மையான நேர்வுகளைக் கையாளுவதற்கான வழக்கமுறைகளால் முன் தொடரப்படும். பொருள் மயக்க நீக்கத்திற்குக் குறிப்பிட்ட நடவடிக்கையை வேண்டும் குறிப்பிட்ட சொல் அல்லது சொற்களின் வகுப்புகள் உள்ளன. ஒரு தனித்தன்மையான நேர்வு ஜெர்மன் சொல் bis ஆகும்; இது துணைநிலை இணைப்புக் கிளவியாக (subordinate conjunction) வரலாம்; சமநிலை இணைப்புக் கிளவியாக (coordinate conjunction) வரலாம்; வினையடையாகவோ முன்னுருபாகவோ (preposition) வரலாம். இச்சொல்லுக்கு வகைப்பாட்டுப் பொருள்மயக்கங்களின் இத்தனித்தன்மையான பிணைப்பு காரணமாக இதைத் தனித்தன்மையான நேர்வாகக் கையாள இயலும். பிற எடுத்துக்காட்டுகள் weder…noch ('neither…nor'), um…willen ('for..sake') , not only…but also போன்ற தொடரா இணைப்புக் கிளவிகள் ஆகும்.

இரண்டாவது துணைச் செயற்பாங்கு INHIBIT என்பதாகும்; சில முறைகளில் இது தனித்தன்மையான நேர்வுகளின் வழக்கங்களை முன் தொடரும். இது வகைப்பாடுகளின் சாத்தியமல்லாத வரிசைமுறைகளைத் தேடி அவற்றை நீக்கும். ஜெர்மன் மொழியில் வரையறை அடை (determiner) + முற்றுவினை என்பது சாத்தியமல்லாதது; எனவே தனியாக வருகையில் முற்றுவினையாக விளக்கிக்கொள்ள இயலும் வரையறை அடையால் பின் தொடரப்படும் ஒரு சொலின் (எ.கா. das Verlangen 'the demand') இந்த அர்த்தம் நீக்கப்படவேண்டும்.

DIHOMஇன் மூன்றாவது முக்கிய செயற்பாங்கு நிகழ்தகவுகள் (probabilities) மற்றும் உடனியைபுத்தகவுளின் (compatibilities) அட்டவணை அடிப்படையில் மீதமுள்ள பொருள் மயக்கங்களின் கனமதிப்புகளின் /முக்கியத்துவங்களின் (weightings) கணக்கீடு ஆகும். நிகழ்தகவுகள் வரிசைமுறையில் வரும் இரு வகைப்பாடுகளின் சாத்தியத்தை அளக்கும்: இது சுத்தமாக ஒரு புள்ளியல் அளவீடாகும்; இது ஒரு தனித் தீர்வாக விளையாமல் தீர்வுகளின் தரவரிசைப்படுத்தப்பட பட்டியலாக அமையும். எடுத்துக்காட்டாக பெயர்-பெயரடை மயக்கம் ஒரு

வரையறை அடையைப் பின் தொடர்ந்தால் அது ஒரு பெயராக இருப்பது கூடுதல் சாத்தியமாகும்; ஆனால் குறைந்த மதிப்பெண்ணுடன் பெயரடையாக வரும் சாத்தியமும் ஏற்றுக்கொள்ளப்படும். உடனியைபுத்தகவும் ('compatibility' ஒத்த அளவீடாகும்; இது பரஸ்பர நிகழ்தகவுகளைக் கணக்கில் எடுக்கும்; அதாவது இரண்டு பொருள் மயக்கமுள்ள சொற்கள் சேர்ந்து வந்தால் ஒன்றின் தீர்வின் நிகழ்தகவு மற்றொன்றின் தீர்வின் உடனியைப்புத்தகவைப் பொறுத்து அமையும். உடனியைப்புத்தகவுப் பரிசோதனை அதிக தூரத்தில் வரும் சொற்களுக்கு இடையிலுள்ள உறவுகளையும் தேடிக்கண்டுபிடிக்கும்: எடுத்துக்காட்டாகச் சார்பு மாற்றுபெயர் (relative pronoun) இருந்தால் அதன் வலப்பக்கம் எங்காவது முற்றுவினை இருக்கவேண்டும். இறுதி நடவடிக்கை சொல் வரிசைமுறைகளுக்கு நிகழ்தகவுகளின் தரவரிசைபடுத்தப்பட ஒழுங்குமுறையின் கணக்கீட்டில் தனிச் சொற்களின் கனமதிப்பைச் சேர்ப்பதாகும்.

எவ்வாறு தோல்வி-மென்பொருள் இயங்குமுறை இந்த அணுகுமுறையை பயன்படுத்திகொள்ள இயலும் என்று பார்ப்பது எளிதாகும்: பின்னர் வரும் தொகுதிகளில் விருப்பப்பட்ட ஒப்புருமொழி பொருள்மயக்கநீக்கம் (homograph disambiguation) தவறானது கண்டுகொண்டால் திருப்பிவந்து அடுத்த தரவரிசைப்படுத்தப்பட தீர்வை முயல்வது ஒப்பீட்டளவில் எளிய விஷயமாகும். இந்த நடவடிக்கைகளை அடிப்படையாகக்கொண்ட மாற்றுமுறை பெரிய அளவிலான தரவுத்தொகுதிப் பகுப்பாய்வில் பயன்படுத்தப்படுவதாகத் தெரியவில்லை (Hutchins and Somer, 1992: 199).

3.8.5.4.**தொடரமைப்புசார் பகுப்பாய்வு**

அடுத்த தொகுதி SEGMENT எனப்படும். அதன் நோக்கம் வாக்கியத்திற்குள் எச்சத்தொடர் எல்லைகளை (clause boundaries) அடையாளங்காண்பதாகும். அதாவது இது வாக்கியத்தை முதன்மை எச்சத்தொடர்கள் (main clause), துணைநிலை எச்சத்தொடர்கள் (subordinate clauses), அடைப்புக்குறிசார் எச்சத்தொடர்கள் (parenthetical clauses) எனக் கூறுபடுத்தும். எச்சத்தொடர் எல்லைகளை நிறுத்தற்குறிகளைக் கண்டுபிடித்து அடையாளம் காணலாம் (ஜெர்மன் மொழியில் இந்த மரபு கண்டிப்பானவை எனவே நம்பகமனதாகும்). முக்கியமான சிக்கல் இணைப்புக்கிளவிகளின் நோக்கவெல்லையை/செயற்பரப்பை (scope) அறிந்துகொள்வதாகும் எ.கா. *und* ('and') என்பது இரண்டு பெயர்களையா இரண்டு பெயர்த்தொடர்களையா இரண்டு வினைத்தொடர்களையா வேறு ஏதாவது இரண்டையா

இணக்கிறது எனத் தீர்மானிப்பது. இந்தத் தொகுதி செல்லுபடி சரிபார்ப்பையும் (validity check) உள்ளடக்கு; இது எச்சத்தொடருக்குள் கட்டாயத் தனிமங்களைத் (obligatory elements) தேடும். எடுத்துக்காட்டாக ஒரு வாக்கியத்தில் முதன்மை எச்சத்தொடர் (main clause) ஒரு வினைமுற்றைக் கொண்டிருக்கவேண்டும்; ஒரு வினையெச்சத்தொடர் (infinitive clause) வினையெச்ச வடிவைக் கொண்டிருக்கவேண்டும் என்பன போன்றன.

வேறுபட்ட கூறுபடுத்தும் வழக்கமுறை (PHRASEG) ஆங்கிலத்திற்கு எழுதப்பட்டுள்ளது; பிரஞ்சுமொழி புற அமைப்பின் பகுப்பாய்வின் மேல் கூடுதல் நெருக்கமாக அடித்தளமிட்டுள்ளது; குறிப்பாக பெயர்கள் மற்றும் வினைமுற்றுகளின் இருப்பிடத்தைப் பொறுத்து அமையும். ஆங்கிலத்திலும் பிரஞ்சுமொழியிலும் நிறுத்தற்குறியிடும் மரபொழுக்கங்கள் ஜெர்மன் மொழி மரபொழுக்கங்களிலிருந்து குறிப்பிடத்தகுந்த அளவு வேறுபடுவதன் காரணமாக இது புகுத்தப்படுத்தப்பட்டுள்ளது.

SEGMENT (அல்லது PHRASEG)இன் வெளியீடு பெயர் குழுக்களையும் வினைக்குழுக்களையும் கையாளும் தொடர்மைப்புசார் பகுப்பாய்வின் இரண்டு நிலைகளுக்கு அனுப்பப்படும். இந்தத் தொகுதிகள் முறையே NOMA (Nominalanalyse) மற்றும் VERBA (Verbalanallyse) என்பன ஆகும்; சூசியின் விளக்கங்கள் இந்தத் தொகுதிகளுக்கு வேறு பெயர்களைக் கொண்டிருக்கின்றது.

NOMA முந்தைய தொகுதிகளால் அடையாளங் காணப்பட்ட கூறுகளின் மேல் இயங்குகின்றது. இது பெயர்த்தொடர்களின் உள் அமைப்புகளைக் கையாளும் இருபது அல்லது அதற்கு மேற்பட்ட துணைச் செயற்பாங்குகளைக் கொண்டுள்ளது. இந்தத் துணை செயற்பாங்குகள் எதிர்ப்பு, எண்கள், பெயரெச்சக் குழுமங்கள், உட்படு அமைப்புகள் போன்ற குறிப்பிட்ட செயல்பாடுகளுக்கு ஒதுக்கப்பட்டுள்ளன. சில 'அடையாளங்கள்' (‘markers’) (நிறுத்தற்குறிகள் அல்லது குறிப்பிட்ட சொற்கள் போன்றன, எ.கா. *for instance* என்பதை எதிர்ப்பின் அடையாளமாகக் கொள்வது) அடிப்படையில் அமையும்; பிற இணைதிறன் உறவுகளை (valency relations) நிறுவுவதற்காக பொருண்மைப் பண்புக்கூறுகள் அடிப்படையில் அமையும். இருப்பினும் முக்கியமான துணைச் செயற்பாங்குகள் முன்னுருபு + வரையறை அடை + பெயர், வரையறை அடை + பெயரடைக் குழும் + பெயர் போன்ற எளிய குழுக்களை அறிவதைக் கையாளுகின்றது. இந்தச் செயல்முறை (கூட்டில் உள்ளது போல் அல்லாமல்) அனுமதிக்கப்பட்ட

வரிசைமுறைகளின் பட்டியல்களின் அடிப்படையில் அமையும்; இது பொருத்தமான வேற்றுமை, பால், எண் இவற்றுடனான முற்றிசைவுகளின் (congruences) அடிப்படையில் அமையும் சாத்தியங்களைக் கொண்டிருக்கும். 'நீளமான பொருத்தம்' (longest match) பிரயோகிக்கப்படும்; அதாவது வகைப்பாடுகளின் நீண்ட வரிசைமுறைகள் முதலில் பரிசோதிக்கப்படும்.

NOMA அதைப்போன்ற வினைக் குழும பகுப்பாய்வின் செயற்பாங்கான VERBA என்பதால் பின் தொடரப்படும். NOMA பெயர் அமைப்புகளின் வேறுபட்ட வகைகளைக் கையாளுவது போன்று VERBAவும் கூட்டுவினைக் குழுக்களைக் கையாளுகின்றது. இது ஜெர்மன் மொழிக்கு முக்கியச் செயல்பாடாகும்; அது ஒப்பீட்டளவில் (எ.கா. பிரஞ்சுமொழியியுடன் ஒப்பிடுகையில்) கொஞ்சம் திரிபுற்ற காலங்களையே கொண்டிருக்கும்; ஆனால் துணவினைகள் மற்றும் வினைநோக்கு வினைகளின் சேர்க்கைகளை அனுமதிக்கும். எ.கா. (6); மொழிபெயர்ப்பு சுட்டிக்காட்டுவது போல் ஆங்கிலத்தில் ஒத்த சிக்கல்கள் உள்ளன.

(6) *Die Aufgabe hätte getan weden sollen*

'The task ought to have been done')

3.8.5.5. **அமைப்புசார் பகுப்பாய்வு**

NOMA மற்றும் VERBA இவற்றின் வெளியீடு பகுப்பாய்வின் அடுத்த நிலைக்கான கட்டுமானப் பொருள்களைத் தரும்; அதாவது KOMA (Komplementanalyse) தொகுதியால் செயல்படுத்தப்படும் அமைப்புசார் அல்லது நிரப்பி (complement) பகுப்பாய்வு. KOMAவின் நோக்கம் ஒவ்வொரு எச்சத்தொடருக்கும் இணை திறனைத் (valency) தீர்மானிப்பதாகும். எடுத்துக்காட்டாக KOMA NOMAவால் அடையாளம் காணப்படும் பெயர் குழுக்களில் எவை வினை எச்சதொடரில் VERBAவால் அடையாளங் காணப்படும் வினைக்கான பங்கேற்பாளர் (அல்லது நிரப்பி) தடங்களை (slots) நிரப்பவேண்டும் எனத் தீர்மானிக்கின்றது. கலவைத்தன்மையான பெயர் குழுக்களின் நேர்வில் அதன் செயல்பாடு உள் இணைதிறன் (internal valency) மற்றும் பங்கேற்பாளர் உறவுகளைத் (argument relations) தீர்மானிப்பதாகும். வினைப் பங்கேற்பாளர் பெயர் குழுக்கள், முன்னுருபுத் தொடர்கள், வினையெச்ச மற்றும் துணைநிலை வினைஎச்சத் தொடர்கள் (subordinate clauses) என்பவைகளை உள்ளடக்கும். வினைகளின் பங்கேற்பாளர்கள் பெயரடை குழுமங்கள், முன்னொட்டுத் தொடர்கள் மற்றும் சிலவகை துணைநிலை எச்சத்தொடர்கள். எ.கா. aஇல் உள்ளது போன்ற daβ துணைநிலை

எச்சத்தொடர்கள் (subordinate clauses) அல்லது b-இல் உள்ளது போன்ற செய்ய-வடிவ எச்சதொடர்கள் (infinitivals) (Hutchins and Somer, 1992: 200).

(7)a. die Idee, daβ er kommen soll 'the idea tht he should come'

b. sein Versperchen, punktlich zu kommen 'his promise to come on time'

KOMA எச்சத்தொடரில் உள்ள முக்கியச் சொல்லுக்குப் பொருத்தமான இணைதிறன் தடங்களை (valency solts) நிரப்புவதை முதலில் முயன்று செயல்படும், எ.கா. முற்றுத் தொடர்களுக்கு எழுவாய் மற்றும் செயப்படுபொருள். இதைத் தொடரியல் வேற்றுமை மற்றும் எண் அடிப்படையில் உடனியைபுத்தகவுளுக்கு (compatibilities) வேண்டி பரிசோதிப்பதாலும் ஓரளவுக்கு எளிய பொருண்மைப் பண்புக்கூறு இணை நிகழ்வுக் கட்டுப்பாட்டுகளுக்கு (cooccurence restrictions) வேண்டி பரிசோதிப்பதாலும் நிறைவேற்றப்படும். மீதமிருக்கும் தனிமங்கள் வினையடைகளாகவோ நேரியைபுகளாகவோ (appositionals) பகுப்பாய்வு செய்யப்படும்; இருப்பினும் உடனியைபுத்தகவுக்கு (compatibility) வேண்டி ஏதாவது பரிசோதனை இருக்கும்.

இந்தத் தொகுதி சொல்நீக்கப்பட்ட அமைப்புகளை (elliptical structures) மீட்டுருவாக்கம் (reconstruct) செய்ய முயலும், எ.கா. (8)a -யிலிருந்து (8)b (Hutchins and Somer, 1992: 201).

(8)a. De Bauer lacht und singt.

b. Der Bauer lacht und Bauer singt.

'The farmer laughs and [the farmer] sings'

இதுபோன்று ஆங்கில KOMA பின்வரும் வாக்கியங்களில் உள்ளது போன்று அக எழுவாய்களையும் (deep subjects) கண்டுபிடிக்க முயலும் (Hutchins and Somer, 1992: 201).

(8)c. John persuaded him to go home.

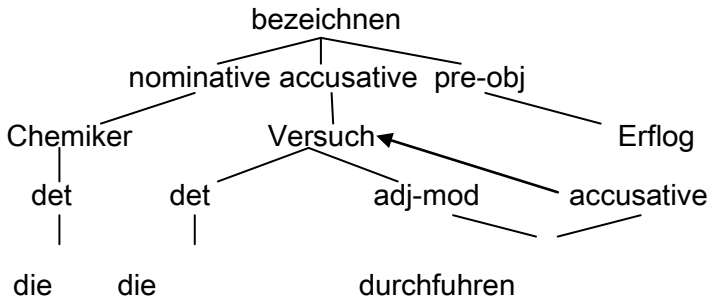
d. John promised him to go home.

இணைதிறன் அடிப்படையில் பகுப்பாய்வு செய்யப்படுவதன் காரணமாக செயப்பாட்டுவினை கட்டுமானங்கள் (passive constructions) பின்வரும் எடுத்துக்காட்டில் KOMAவிலிருந்து வெளியீட்டின் எடுத்துக்காட்டு (9) எடுத்துக்காட்டுவது போல் செய்வினைக் கட்டுமானங்களாக உருப்படுத்தம் செய்யப்படும்.

(9) a. *Die durcheführten Versuche werden von dem Chemiker als*

‘The experiments carried out are declared a success by the chemist’

(9)b



3.8.5.6. **பொருண்மை மயக்கநீக்கம்**

பகுப்பாய்வு கட்டத்தில் இறுதித் தொகுதி SEDAM (Semantische Disambiguierung) ஆகும். இதன் செயல்முறைகள் முக்கியமாக 'பொருண்மை அகராதிகளில்' உள்ள பெயர்களுக்கும் சில மாற்றுப்பெயர்களுக்கும் ஒதுக்கப்பட்டுள்ள பொருண்மைப் பண்புக்கூறுகள் மற்றும் பதிவுகளுடன் இணைக்கப்பட்டுள்ள பொருண்மை விதிகளின் வேறுபட்ட வகைகளின் அடிப்படையில் அமையும். பொருண்மைக் கூறுகள் இரு வகைப்படும்: (மனிதன், அருவம், விலங்கினம் போன்ற) உலகப் பொதுமைகளாகக் கருதப்பட்டவைகளின் எல்லைக்குட்பட்ட தொகுப்பிலிருந்து எடுக்கப்பட்டவை; (புவியல் இடம், ஊர்தி, விலங்கு, தாவரம் போன்ற) தொடக்கத்தில் பெயர்களுக்கு வேண்டி உருவாக்கப்பட்ட பண்புக்கூறுகளின் குறிப்பிட்ட மொழிக்குரிய தொகுப்புகளிருந்து எடுக்கப்பட்டவை. முதல் தொகுப்பு படிநிலை அமைப்பில் அமையும்; இதனால் துணைப் பண்புக்கூறுகள் அகராதியில் புகுத்தப்படவில்லை; பொருத்தமுள்ளதாக தனியக்கமாகமாக உருவாக்கப்படுகின்றது. பதிவுகளுடன் இணைக்கப்பட்ட விதித் தொகுப்புகள் அமைப்புகளில் மாற்றங்களுக்கு, நீக்கலுக்கு, சொருகல்களுக்கு வேண்டி அழைக்கப்படும்; அவை சூழலின் தொடரியல் அல்லது பொருண்மையியல் தனிக்குறிப்பீடுகளால் கட்டுப்படுத்தப்படும்.

SEDAMஇன் நோக்கம் பொருண்மை அடிப்படையில் மயக்கம் உள்ளதாகவும் தெளிவற்றதாகவும் உள்ள தொடரியல் அமைப்புகளுக்குப் பொருள்கோளை ஒதுக்குவதாகும்.

எடுத்துக்காட்டாக, ஒரு செயற்பாங்கைக் குறிப்பிடும் ஒரு பெயரை வெளிப்படுத்தப்படாத பல எண்ணிக்கையிலான வழிகளில் அடைசெய்ய இயலும்: ஒரு அடைசெய்யும் பெயர் செயற்பாங்கின் செயலியையோ (வினைமுதல்) (எ.கா. human understanding) செயப்படுபொருளையோ (satellite launching) முறையையோ (computer stimulation) இடத்தையோ (inner-city deprivation) இவை போன்றவற்றையோ குறிப்பிடலாம். இம்மாதிரியான பொருண்மை மயக்கங்கள் பொருண்மை விருப்பங்களை வெளிப்படுத்தும் விதிகளின் பிரயோகத்தால் தீர்க்கப்படும்.

SEDAM முயலுகின்ற மற்றொருவகை பொருண்மை மயக்கம் ஒரே வகைபாட்டின் ஒப்புருமொழிகளின் (homographs) வேறுபாடு ஆகும். அதாவது இதுவரை சொற்களை ஒரு தனி தொடரியல் வகைப்பாட்டிற்கு ஒதுக்க இயலும்; ஆனால் அவைகளுக்கு இரண்டோ அதற்கு மேலோ தனித்துவமான அர்த்தங்களும் பயன்பாடுகளும் இருக்கும். எடுத்துக்காட்டாக trade என்பது பணியையோ (occupation) தொழிலையோ (profession) பொதுவான வணிகத்தையோ குறிப்பிடலாம். முதல் அர்த்தத்தை உடைமை மாற்றுப்பெயர் (his trade) இருப்பதாலோ முன்னுருபுக்கு முன்னால் வருவதாலோ (by trade) பிரித்தெடுக்கலாம்; பெயரானது பன்மை என்றால் இரண்டாவது அர்த்தத்தை ஒதுக்கி வைக்கலாம். பெரும்பாலான அகராதிகள் பொருண்மைப் பண்புக்கூறுகளுக்குப் பிரயோகிக்கப்படும் இணை நிகழ்வு (சேர்ந்துவருகை) கட்டுத்திட்டங்கள் (co-occurrence restrictions) அடிப்படையில் அமைந்துள்ளன, எ.கா. raise என்பதன் பல்வேறு பயன்பாடுகளைக் ('cultivate', 'express', 'produce', 'bring up') அடையாளங் கண்டுகொள்வது. சந்தேகத்திற்கிடமின்றி இந்த வேறுபாடுகள் ஓரளவுக்குக் குறிப்பிட்ட மொழிகளுக்குச் சொல்சார் மாற்றலுக்குத் தேவையான தகவல்களைத் தருவதன் தேவையால் ஊக்குவிக்கப்பட்டுள்ளது; சூசி ஆய்வாளர்கள் இவற்றை எந்த அளவுக்குக் கூடுதல் விரிவாகப் பயன்படுத்தக் கருதுகிறார்கள் என்பது தெளிவாகத் தெரியவில்லை (Hutchins and Somer, 1992: 202).

மாற்றல் சிக்கல்களை நேரிடுவதற்காகவும் 'பகுதி-நிலைத்த' (semi-fixed) வெளிப்பாடுகளை அடையாளங்காண்பதற்காக நடைமுறைகள் SEDAM-இல் உட்படுத்தப்பட்டுள்ளது. எ.கா. take into consideration என்ற தொடரின் அமைப்புகளில் பிற கூறுகள் குறுக்கிடலாம் (Hutchins and Somer, 1992: 202).

(10)The commission took the plan into consideration.

அடிப்படையான செயல்பாடு commission என்பதை எழுவாயாகவும் plan என்பதைச் செயப்படுபொருளாகவும் கொண்டு வினை + பின்னுருப்தொடரை ஒரே அலகாகப் பொருகோள் செய்வதாகும்.

பின்னுருபுகள் தனித்தன்மையான நடைமுறையால் கையாளப்படுகின்றன; இது பங்கேற்பாளர்களின் இணைதிறன் உறவுகளையும் பொருண்மைப் பண்புக்கூறுகளையும் பரிசோதிக்கும்; மற்றும் 'செயற்கைச் சொற்களை' (artificial words) உருவாக்கும்; இதன் விளைவாகத் தொடர்ச்சியான மாற்றல் செயல்பாடுகளில் மாற்றப்படாமல் விடப்படும் இடைமொழித் தனிமங்கள் உருவாகும்.

3.8.5.7. **மாற்றலும் உருவாக்கமும்**

TRANSFER தொடர்ச்சியான செயற்பாங்குகள் ஆகும். இது இருமொழிய அகராதியைப் பயன்படுத்தி மூலமொழிச் சொல்வடிவுகளை அவற்றின் இலக்குமொழி நிகரன்களால் இடம்பெயர்க்கின்றது. சாதாரணமாக, தொடர்புடைய சார்பு அமைப்பு பாதுகாக்கப்படும்; இருப்பினும் விதிவிலக்காக இது TRANSFERஆல் மாற்றப்படலாம். இந்தத் தொகுதி ஒரு துணைச் செயற்பாங்கை உள்ளடக்கும்; இது WOBUSU அறியப்படாதன என அடையாளங்காணும் சொற்களை மொழிபெயர்க்க முயலும். சொல் அலகுகளின் எளிய மொழிபெயர்ப்பு TRANSFERஇன் முக்கியப் பங்களிப்பாகும்; இது சுற்றியிருக்கிற சூழலைப் பார்க்காமல் சொல்லை அடுத்து சொல் (word-by-word) என்ற அடிப்படையில் செய்யப்படுகின்றது. TRANSFERக்குள் சில சிறப்பு நோக்கமுள்ள துணைத் தொகுதிகள் உள்ளன; அவை எதிர்மறையை மொழிபெயர்த்தல், பெயர் குழுக்களுக்குள் புற வேற்றுமை உறவுகளைக் கையாளுதல் போன்ற குறிப்பிட்ட நேர்வுகளைக் கையாளும். சில மொழி இணைகளுக்குக் கூடுதலான TRANSFER தொகுதிகள் தேவைப்படும்; எடுத்துக்காட்டாக ரஷ்யமொழியை மூலமொழியாகக் கொண்டு மொழிபெயர்க்கும் போது வரையறை அடைகளின் பற்றாக்குறையைக் கையாளவேண்டும்.

TRANSFERஇன் வெளியீட்டின் அடிப்படையில் இலக்குமொழி வாக்கியங்களை உருவாக்குவது SYNTHESEயின் செயல்பாடு ஆகும். இது மூன்று துணைச் செயற்பாங்குகளை கொண்டிருக்கின்றது; இவை பொருண்மைசார், தொடரியல்சார், உருபனியல் சார் உருவாக்கத்தைக் கையாளும்.

பொருத்தமான இடத்தில் மரபுச்சொற்களையோ (idioms) அல்லது பகுதி-நிலைத்த கட்டுமானங்களையோ ('semi-fixed constructions) உருவாக்குவது அல்லது வினைகளுக்கும் பெயர்களுக்கும் சரியான பின்னுருபுகளை உருவாக்குவது (எ.கா. புவியல் பெயர்களுடன் to-ஐ மொழிபெயர்ப்பது: *in die Schweiz, zu Berlin, nach Europa*) மற்றும் செயற்கைச் சொற்களை ('artificial words') இலக்கு மொழிச் சொற்களாக மொழிபெயர்ப்பது ஆகியன SEMSYN (Semantische Synthese) என்பதன் செயல்பாடாகும். செயற்கைச் சொற்கள் பகுப்பாய்வின் போது உருவாக்கப்படுகின்ற சொற்கள் ஆகும் மற்றும் பொருண்மை மயக்கம் நீக்கப்பட்ட ஒப்புருமொழி மற்றும் தொடர்ச்சியற்ற பன்மடங்குச்சொற்கள் (multi-word lexical items) (*to fram* என்பதன் மொழிபெயர்ப்பாக *Landwirtschat betreiben*; *away from* என்பது போன்ற கூட்டு முன்னுருபுகள், *take down* போன்ற ஆங்கில தொடர் வினைகள் போன்றன).

SYNSYN (Syntaktische Synthese) SEMSYNஇன் வெளியீட்டை எடுத்துக்கொண்டு சொற்களின் (அதாவது சொல் பகுதிகளின்) வரிசையை உருவாக்குகின்றது; இது உருபனியல் உருவாக்கத்திற்கான தொடர்புடைய உருபனியல் தகவல்களைக் கொண்டிருக்கும். சரியான புற சொல் வரிசையையும், பெயர்களின் புற வேற்றுமை உருபுகளையும், சம்பந்தப்பட்ட மூலமொழியைப் பொறுத்து பல உடன்பாட்டு நிகழ்வுகளையும் (எ.கா. ஜெர்மன் மொழியின் கொடைவேற்றுமை பன்மை உருபு, பல மொழிகளில் எழுவாய்-வினை உடன்பாடுகள், பிரஞ்சுமொழி பெயர் குழுமங்களில் எண்-பால் உடன்பாடு) தீர்மானிப்பது இந்தத் தொகுதியின் பகுதிச் செயல்பாடு ஆகும்.

SYNSYNக்குள் (Syntaktische Synthese) பல ஆர்வமூட்டும் துணைச் செயற்பாங்குகள் உள்ளன. SGKOMP எனப்படும் துணைச்செயற்பாங்கு கூட்டுக்களையும் ஆக்கங்களையும் கையாளுகின்றது. எடுத்துக்காட்டாக, *système de transduction* ('translating system') என்ற பிரஞ்சுமொழித் தொடரை ஜெர்மன் மொழியில் மொழிபெயர்த்தால் அது ஒரு கூட்டுப் பெயரை உருவாக்கும் System von [or für] Ubersetzung என்பதற்குப் பதிலாக *Übersetzungssystem* என்பதை உருவாக்கும். SNOADJ என்ற மற்றொரு தொகுதி சில அடை-பெயர் பிணைப்புக்குப் பல சாத்தியமான மெய்படுத்தங்களில் எதைத் தெரிந்தெடுக்கவேண்டும் என்பதைத் தீர்மானிக்கும். எ.கா. *Israeli lemons* என்பதா அல்லது *lemons from Israel* என்பதா வெளியீடாகத் தெரிந்தெடுக்க வேண்டும்.

SYNSYNஇன் மிகக் கலவைத்தன்மையான/சிக்கலான துணைச் செயற்பாங்குகளில் ஒன்று SPORE ஆகும்; இது மாற்றுப்பெயர்கள் மற்றும் பெயரடைகள் இவற்றின் உருவாக்கத்தைக் கையாளுகின்றது. எடுத்துக்காட்டாக, ஜெர்மன் மொழியில் மாற்றுப் பெயர்கள் அவற்றின் முன்வரும் பெயர்களுடன் இலக்கணப் பாலால் (grammatical gender) உடன்படவேண்டும். கீழ்வரும் வாக்கியத்தில் *it* என்பது *file* என்பதைக் குறிப்பிடும் (அது ஜெர்மன் மொழியில் *Datei* ஆகும்); *system* என்பதைக் குறிப்பிடாது (அது அஃறிணை/ஒன்றன்பால் ஆகும்); எனவே அது *sie* என மொழிபெயர்க்கப்படவேண்டும்; *es* என மொழிபெயர்க்கப்படக் கூடாது (Hutchins and Somer, 1992: 203).

(11) If the system deletes a file, the user can recover it by…

இது போன்ற சிக்கல் ஜெர்மன் மொழியின் உடைமை பெயரடைகளில் உள்ளது; இதுவும் முன்னர் கூறியது போல் அதன் முற்சுட்டுகளுடன் இலக்கணப் பால் உடன்பாட்டைக் காட்டவேண்டும்; இது பெரும்பாலும் ஒன்றன்பாலுடன்/அஃறிணையுடன் முரண்படும். பின் வரும் எடுத்துக்காட்டு இதை உணர்த்தும் (Hutchins and Somer, 1992: 203).

(12) Die Gemeinschaft und ihre Mitglieder

'the community and its members'

Lit: The community and her membrs

வாக்கிய எல்லையைக் கடந்த முற்சுட்டைக் கையாளவும் முயற்சி மேற்கொள்ளப்பட்டது; இது சாத்தியமான முன்வரும் சொற்களை அவை வரும் இடங்கள் மாறும் அவற்றின் பங்குமுறைகளின் (roles) படி மதிப்பீடு செய்யும் ஒரு ஒழுங்குமுறையைப் பயன்பத்தி செய்யப்படும்.

MORSYN (Morphologische Synthese) என்ற தொதியைக் கொண்டு மொழிபெயர்ப்புச் செயற்பாங்கு முடிவுறும். பகுதி + உருபனியல் பண்புக்கூறு அலகுகளை இலக்கு மொழிப் பனுவல் கோர்வைகளாக மாற்றுவது இதன் நோக்கம் ஆகும். இது சிலவேளைகளில் ஒப்பீட்டளவில் நேரடியானதாக இருக்கும்; இருப்பினும் சில மொழிக்ளில் (குறிப்பாக ஜெர்மன் மொழியில்) உருபனியல் வடிவுகள் அவற்றைச் சுற்றியுள்ள சூழலை அதிக அளவில் சார்ந்திருக்கும். ஜெர்மன் மொழியின் பெயரடை ஈற்றுருபுகள் (adjective endings) பெயரடை விசேடணம் செய்யும் பெயரின் வேற்றுமையையும் பாலையும் சார்ந்து அமைவதோடு எந்த வகையிலான வரையறை அடையைப் பெயர் குழுமம் கொண்டிருக்கின்றது என்பதைப் பொறுத்தும் அமையும், எ.கா. *ein*

guter Mann, der gute Mann. இது போன்று ஆங்கிலமும் பிரஞ்சுமொழியும் உருபனியல் பிறழ்ச்சியைக் கொண்டுள்ளன; இது இலக்கணம் சார்ந்ததாகும்; சூழல் சார்ந்ததல்ல. எ.கா. பிரஞ்சுமொழி *de le*-ஐ *du*-ஆல் இடம் பெயர்ப்பது; உயிர்கள் முன்பு அசை கெடுதல் *l'*, *d'*; ஆங்கில *a* அல்லது *an* என்பதன் தெரிவு). இறுதியாக MORSYN தலைப்பெழுத்தாக்கம், நிறுத்தற்குறியிடுதல் என்பனவற்றையும் கையாளுகின்றது; இரண்டும் ஜெர்மன் மொழி வெளியீட்டுக்கு முக்கியமானவை ஆகும்.

3.8.7. **முடிவுரை**

முக்கியமான அம்சங்களில் சூசி மாற்றல் அடிப்படையிலான ஒழுங்கமைப்புக்கு எடுத்துக்காட்டாகும்; ஆனால் சில அம்சங்களில் இது ஒரு கலப்பு ஒழுங்குமுறையாகும். கணினித்துவக் கண்ணோட்டத்தில் இது அடிப்படையில் முதல் தலைமுறை வகை நிர்மாணத்தை வெளிக்காட்டுகின்றது; இது சிஸ்டிரானில் காண்பதை ஒக்கும். மொழியியல் நடைமுறைகள் உயர் நிலை விதி எழுதும் வடிவவாதத்தில் (rule-writing formalism) எழுதப்படவில்லை; கீழ் நிலை நிரலாக்க மொழி ஃபோர்ட்ரானில் நேரடியாக எழுதப்பட்டுள்ளன. சூசியின் பலவீனம் நிரலாக்க சூழலும் உடனிருக்கிற பழமையான தரவு அமைப்பும் மற்றும் எளிமையான செயலாக்கமும் ஆகும்; குறிப்பாக தனியான கிளையற்ற தொகுதிகளின் வரிசை ஆகும் (Hutchins and Somer, 1992: 203).

சூசியின் கலப்பு நேரடிப் பரம்பரையின் ஒரு சிறப்பியல்பு இதன் முதல் நிலையின் பகுப்பவில் மொழியியல் அல்லாத முறைகளை வலியுறுத்துவதாகும். SEGMENT செயல்முறைகள் வரை, மரபு (கணினி மொழியியல்) அர்த்தத்தில் பகுப்பாய்வு (parsing) என்று கூறத்தக்க ஒன்றும் இல்லை. பல சிக்கல்கள் தற்காலிக விதிகள் அடிப்படையில் தீர்க்கப்படுகின்றன.

பகுப்பாய்வு செயல்முறைகள் முற்றிலும் ஒருமொழிசார்ந்ததல்ல. இலக்கு மொழி சார்ந்த பகுப்பாய்வு குறிப்பாக SEMSYNஇல் காணப்படுகின்றது. பொதுவாக சூசியின் மொழிபெயர்ப்பின் தரத்தைப் பிற சமகால ஒழுங்குமுறைகளுடன் ஒப்பிட இயலும். சூசி பல நடைமுறை பரிசோதனைகளில் பயன்படுத்தப்பட்டுள்ளது. ஒரு கூட்டு ஆய்வுத்திட்டத்தில் ஜெர்மன் மொழிக்கும் ஜப்பானிய மொழிக்கும் இடையிலுள்ள ஆவணங்களின் தலைப்புகளை மொழிபெயர்க்க சூசி கொயோட்டோ பல்கலைக்கழகத்தின் TITRAN என்ற ஒழுங்குமுறையுடன் இணைக்கப்பட்டது.

பகுதி வெற்றி கிடைத்தபோதிலும், சூசியின் மேம்பாட்டாளர்கள் அசல் ஒழுங்குமுறையின் வரம்புகளை அறிந்துகொண்டனர்; எனவே புதிய ஆய்வுத்திட்டமான SUSY-II என்ற ஆய்வுத்திட்டத்தை தொடங்கி செயல்படத் தொடங்கினர். இதன் நோக்கம் மொழியியல் விதி எழுதும் வடிவவாதத்தை உருவாக்குவது, பகுப்பாய்வு நிலைகளின் எளிமையாக்கம் (மூன்று அடிப்படை செயற்பாங்குகளாகக் குறைத்தல்), அமைப்புப் பொருண்மை மயக்கத்தை கூடுதல் திருப்தியாகக் கையாளுதல், விருப்ப விதிப் பிரயோகத்தைப் (preferential rule application) புகுத்துதல்.

3.9. மெட்டெயோ (MÉTÉO)

Météo ஒழுங்குமுறை TAUM குழுவால் காலநிலை செய்திகளை ஆங்கிலத்திலிருந்து பிரஞ்சு மொழிக்கு மொழிபெயர்ப்பதற்காக மாண்ட்ரியலில் உருவாக்கப்படதாகும். இந்த ஒழுங்குமுறை 1976-இல் நிறுவப்பட்டது. இதன் வெற்றி, காலநிலை முன்னறிவுப்புகள் என்ற துணைநிலைமொழிக்கு கட்டுப்படுத்தபட்டதன் காரணமாக முழுமையையும் துல்லியத்தையும் பெறலாம் என்பதன் அடிப்பையில் அமைந்ததாகும்.

TAUM-MÉTÉO உண்மையிலேயே உலகின் முழுமையான தானியங்கு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைக்கு ஒரே எடுத்துக்காட்டாகும் (Slocum, 1985). TAUM தொழில் நுட்பத்தின் பக்கவிளைவாக உருவாக்கப்பட்ட இவ்வியந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை 1977-இல் கனடிய வானிலை ஆய்வு மையத்தின் (Canadian Meterological Centre (CMC)) தேசீய அளவிலான தகவல்பரிமாற்ற பிணையத்தில் முழுவதுமாக ஒருங்கிணைக்கப்பட்டது. MÉTÉO ஆங்கில வானிலை அறிக்கைகளை நேரடியாக மொழிபெயர்த்து தானியக்கமாக தகவல்தொடர்பு பிணையத்திற்கு அனுபித்தந்தது. பிழைகளைக் கண்டுபிடித்துத் திருத்தப் பின்திருத்திகளைச் சாராமல் MÉTÉO அதனுடைய தவறுகளைத் தானே கண்டுபிடித்துத் தவறான உள்ளீட்டை மனிதத் திருத்திகளுக்கு அனுப்பியது; சரியானது என்று MÉTÉO-ஆல் கணிக்கப்பட்ட வெளியீடு மனிதத் தலையீடு இன்றி அனுப்பப்பபட்டது.

TAUM-MÉTÉO வடிவமைப்பு, உருவாக்கம், மேம்படுத்தல் என்ற எல்லா கட்டங்களிலும் மொழிபெயர்ப்பாளர்களை உட்படுத்திய முதல் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை என்று கூறவியலும். MÉTÉO-இன் உள்ளீட்டின் மீதான கட்டுப்பாடுகள் திட்டம் தொடங்குவதற்கு

முன்னரே இருந்தபடியால் சிஸ்ட்ரானின் சிறப்புப்பண்புகளுடன் இணைக்கப்பட்ட கட்டுப்பாடுகளைச் சார்ந்திருக்கிற ஜெராக்ஸ் சிஸ்ட்ரான் ஒழுங்குகுமுறையுடன் MÉTÉO ஒழுங்குமுறையைக் குழும இயலாது. ஆனால் MÉTÉO-ஐ விரிவாக்கம் செய்ய இயலாது.

MÉTÉO நிறுவலின் கூடுதல் பக்கவிளைவு, சிளம்சி மொழிபெயர்ப்பாளர்கள் MÉTÉOஇன் செயல்பாட்டுத் தீர்மானங்களை நம்பத்தொடங்கிய உடன் MÉTÉO-வுக்கு முன்பிருந்த மொழிபெயர்ப்பாளர் எண்ணிக்கை விகிதம் 6 மாதங்களிலிருந்து பல ஆண்டுகளாக உயர்ந்தது. MÉTÉO-இன் உள்ளீடு நாளொன்றுக்கு 24,000 சொற்களை அல்லது ஆண்டொன்றுக்கு 8.5 மில்லியன் சொற்களைக் கொண்டிருந்தது. இதில் 90-95%-ஐ சரியாக மொழிபெயர்த்தது; பிறவற்றை, 5-10%-ஐ சிளம்சி மனித மொழிபெயர்ப்பாளர்களுக்கு அனுப்பித்தந்தது. இந்த "பகுப்பாய்வு தோல்விகளை" தகவல்தொடர்பு இரைச்சல் (communicatios noise), எழுத்துப் பிழைகள், அல்லது அகராதியில் சொற்கள் இல்லாமையுடன் காரணப்படுத்தலாம்; சில தோல்விகளுக்கு இந்த ஒழுங்குமுறை சில மொழி கட்டுமாங்களைக் கையாளுவதற்கு இயலாமையைக் காரணமாகக் கூறலாம். 1981-இல் MÉTÉO-இன் கோட்பாட்டு அடிப்படையின் உள்ளமைக்கப்பட்ட வரம்புகள் அடையப்பெற்றது; மேலும் மேம்படுத்துவது சிக்கனமாக இருக்காது என்ற நிலையை எய்தியது.

கீழே ஆங்கிலம்-பிரஞ்சு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் அசல் Q-ஒழுங்குமுறைகளின் செயல் படுத்தம் பற்றி விளக்கப்படும் (Hutcins and Somers 1992).

3.9.1. வரலாற்றுப் பின்னணி

மாண்ட்ரியல் பல்கலைக்கழகத்தில் 1965இல் CETEDOL (Centre d'Études pour le Traitement Automatique des Données Linguistiques 'Centre for Automatic Linguistic Data Processing Studies') என்பதன் உருவாக்கத்தால் கை ரோண்டியோ (Guy Rondeau) இயக்கத்தின் கீழ் இயற்கைமொழி ஆய்வு ஆராய்ச்சி தொடங்கியது. இச்சமயத்தில் மொழிபெயர்ப்பு சேவையின் தேவை கணிசமான அளவு வளர்ந்தது. கனடா அரசு இருமொழியக் கொள்கையை அறிமுகப்படுத்தியது; இதன் காரணமாக அரசாங்க ஆவணங்கள் யாவும் ஆங்கிலம் மற்றும் பிரஞ்சு என்ற இருமொழிகளிலும் இருக்கவேண்டிய கட்டாயம் ஏற்பட்டது. கனடா தேசிய ஆராய்ச்சி மன்றம் (Candian National Research Council) இயந்திர மொழிபெயர்ப்பின் நிதிநல்கையைத் தொடங்கியது. 1968க்கும் 1971க்கும் இடையில் மையத்தின் கவனக்குவிப்பு இயந்திர மொழிபெயர்ப்பு ஆக மாறியது மற்றும் அந்தக் குழு TAUM (Traduction Automatique de

l'Université de Montréal) என மறுபெயரிடப்பட்டது. அந்தக் குழு மாற்றல் அடிப்படை அணுகுமுறையைப் பின்பற்றி மற்றும் அலைன் கோல்மரவுர் (Alain Colmerauer)–ஆல் உருவாக்கப்பட்ட 'Q-ஒழுங்குமுறைகள்' ('Q-systems') மென்பொருளில் எழுதப்பட்ட ஒரு மூலமுன்மாதிரி ஆங்கில-பிரஞ்சு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை (Prototype English to French system) உருவாக்கியது. முன்னரே TAUM மொழிபெயர்ப்பில் பல பொருண்மையியல் சிக்கல்களைத் தீர்வுசெய்ய துணைமொழி அணுகுமுறையின் (sublanguage approach) நன்மைகளை உணர்ந்திருந்தனர் மற்றும் 1975இல் பொது வானிலை முன்னறிவிப்புகளை மொழிபெயர்க்க ஒரு ஒழுங்குமுறையை உருவாக்க ஒரு ஒப்பந்தத்தைப் பெற்றது.

இந்தச் செயல்பாடு இயந்திர மொழிபெயர்ப்புக்குப் பல வழிகளில் ஏற்றதாகும். வானிலை அறிக்கைகள் மனச்சலிப்பு தருவதும் திரும்பத்திரும்பக் கூறுவதும் ஆகும். வேலைத் திருப்தி மிகக் குறைவாகும் மற்றும் நாடுமுழுவதும் இருமொழிய வானிலை அறிக்கை கிடைக்க கனடா அரசு தீர்மானித்த போது மொழிபெயர்ப்பு ஊழியர்கள் வரவு அதிகமாக இருந்தது. TAUM கனடா வெளியுறவுத்துறை செயலாளரின் மொழிபெயர்ப்புப் பணியகம் (Translation Bureau of the Canadian Secretary of State) ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கக் கேட்டுக்கொள்ளப்பட்டது. மூலமுன்வகை (prototype) 1976இல் நிரூபிக்கப்பட்டது. மெடெயோ மே 1977இல் முழுநேர செயல்பாட்டைத் தொடங்கியது. அன்றிலிருந்து டோவலிலில் உள்ள மெடெயோவால் மொழிபெயர்க்கப்பட்ட வானிலை முன்னறிவிப்பு கனடிய வானிலை மையத்தால் நாள்தோறும் ஒலிபரப்பட்டது. 1984 அக்டோபரில் புதிய பதிப்பு மெட்டெயோ-2 மைக்ரோ கணினிகளில் செயல்பாட்டிற்காக நிறுவப்பட்டது. மெட்டெயோ-2 ஜான் சாண்டியோக்ஸ் ஆலோசகர்கள் இணைக்கப்பட்ட நிறுவனம் (John Chandioux Consultants Inc.) உருவாக்கிய GramR என்ற நிரலாக்க மொழியால் எழுத்தப்பட்டது (சாண்டியோக்ஸ் அசல் ஒழுங்குமுறையின் முதன்மை வடிவமைப்பாளர் ஆவார்). Q-ஒழுங்குமுறைகளின் மெயின்ஃப்ரேம் பதிப்பு பயன்படுத்தபடவில்லை: மெட்டெயோ-2 விரைவானதாகவும் நம்பகமானதாகவும் செலவு குறைந்ததாகவும் நிரூபிக்கப்பட்டுள்ளது. கியூபெக் வானிலை அலுவலத்தால் வெளியிடப்பட்ட செய்தி அறிக்கைகளின் ஆங்கில-பிரஞ்சு மொழிபெயர்ப்புக்கு உரிய MÉTÉO ஒழுங்குமுறையை நிறுவுதல் 1989இல் ஏற்பட்ட மேலும் வளர்ச்சி ஆகும்.

கீழே ஆங்கில-பிரஞ்சு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் Q-ஒழுங்குமுறைகள் செயல்படுத்தம் விளக்கப்பட்டுள்ளது.

3.9.2. மொழிபெயர்ப்புச் சூழல்: உள்ளீடு/இடுபொருள், முன்னாய்வு மற்றும் பின் திருத்தியமைத்தல்

கனடிய வானிலை ஆய்வு மையத்தின் (Canadian Meteorological Center) Météo மொழிபெயர்ப்புத் திட்டம் பெரிய ஒழுங்குமுறையில் உள்ளடக்கப்பட்டுள்ளது; இது தகவல்தொடர்பு வலை அமைப்பில் இருந்து அறிக்கைகளைப் பெறும்; தரவை முன்னாய்வு செய்யும்; மொழிபெயர்க்க இயலாத விஷயங்களை மனிதத் தொகுப்பாளர்களுக்கு அனுப்பும்; Météo வெளியீட்டை மீள்வடிவமாக்கம் செய்யும்; இறுதிப் பதிப்பை வலை அமைப்புக்குக் கடத்தி அனுப்பும். தற்போது இந்த ஒழுங்குமுறை 37000 சொற்களை ஒவ்வொரு நாளும் 90% துல்லியத்துடன் மொழிபெயர்க்கும் (Hutcins and Somers 1992: 208).

வானிலை அறிக்கைகள் படம் 1இல் எடுத்துக்காட்டப்பட்டுள்ளது போல் நிலைபெறு/தரமான வடிவமைப்புகளில் பெறப்படுகின்றது: குறியமாக்கப்பட்ட தலைப்பு, அறிவிப்பு வெளியீட்டின் மூலத்தின் அறிக்கை, செய்தி அறிக்கை பயன்படக்கூடிய இடங்களின் பட்டியல், முன்னறிவிப்பு, இறுதியில் ஒரு முடிப்பான். இந்த வடிவமைப்பு கண்டிப்பாகப் பின்பற்றப்படுகின்றது. அறிவிப்பு வெளியீடுகளின் சொற்றொகை இதுபோல் திடமானதாகும் மற்றும் தலைப்புகள், இடப்பெயர்கள், வானிலை நிலைமைகள் இவற்றின் ஒரு தொகுப்புத் தொடர்களுக்குக் கட்டுப்படுத்தப்பட்டுள்ளதால் ஊகிக்கக்கூடியதாகும். இதன் அர்த்தம், உள்ளீடு செய்யப்பட்ட பனுவலை ஒலிபரப்புப் பிழையால் ஏற்பட்டுள்ள 'அறியப்படாத சொற்களுக்காக' பரிசோதிக்க இயலும். கிட்டத்தட்ட இவைதான் மனித மொழிபெயர்ப்பாளர்களால் திருத்தப்படுவதற்காக அனுப்பப்படும் அறிக்கைகளில் விளையும் பிழைகளாகும்.

முதல் செயன்மை மொழிபெயர்ப்பு அலகுகளை அடையாளம் காணல் மற்றும் அவற்றை Q-ஒழுங்குமுறை வடிவமைப்புக்கு மாற்றுதல் இவற்றை உள்ளடக்கும்; தனிச் சொற்களும் நிறுத்தற்குறிகளும் '+' ஆல் பிரிக்கப்பட்டிருக்கும்; தொடர்களின் தொடக்கமும் முடிவும் '-01' மற்றும் '-02' இவற்றால் பிரிக்கப்பட்டிருக்கும். மேலும் இது ஒரு அடையாளங்காட்டியால் குறிக்கப்பட்டிருக்கும்; தேவையென்றால் இதனால் வாங்கி (receiver) தொடரின் மூலத்தை இடங்காண இயலும். இந்த அடையாளங்காட்டிகள் மூலத்தைக் கொண்டிருக்கும், எ.கா. Toronto,

மற்றும் ஒரு இயங்கும் எண் (running number). படம் 1இல் அறிவிப்பு வெளியீடுக்கு வெளியீடு படம் 2இல் உள்ளது போல் இருக்கும்.

படம் 1: அறிவித்தபடி வானிலை அறிக்கை

FPCN11 CYYZ 311630

FORECASTS FOR ONTARIO ISSUED BY ENVIRONMENT CANAD AT 11.30 AM EST WEDNESDAY MARCH 31ST 1976 FOR TODAY AND THURSDAY

METRO TORONTO

WINDSOR.

CLOUDY WITH A CHANCE OF SHOWERS TODAY AND THURSDAY

LOW TONIGHT 4. HIGH THRUSDAY 10

OUTLOOKS FOR FRIDAY … SUNNU

END

பட்டம் 2: வடிவாமாக்கப்பட்ட வானிலை அறிக்கை

-01-$(TORONTO,2) + FORECAST + FOR + ONTARIO + ISSUED+BY + ENVIRONMENT CANADA + AT + 11 + H + 30 + AM + EST + WEDNESDAY + MARCH + 31ST + 1976 + FOR + TODAY + AND + THRUSDAY + . -02-/

-01- $(TORONTO,3) + METRO + TORONTO + , WINDSOR + . -02-/

-01- $(TORONTO,4)+ CLOUDY + WITH + A + CHANCE + OF + SHOWERS + TODAY + AND + THRUSDAY + . -02-/

-01- $(TORONTO,5) + LOW + TONIGHT + 4 + . -02-/

-01- $(TORONTO,6) + HIGH + THRUSDAY + 10 . -02-/

-01- $(TORONTO,7) + OUTLOOK + FOR +FRIDAY + = + SUNNY + .02-/

எடுத்துக்காட்டு காட்டுவது போல், முன்பகுப்பாய்வு சில முறைப்படுத்தலை உட்படுத்தும்; அசல்11.30 என்பது 11 + H + 30 என்று மாறும்; இது சுருக்கத்தின் விரிவாக்கத்தையும் உட்படுத்தும்; எடுத்துக்காட்டாக kmh → kilometers per hour, Jan → January, BC → British Columbia போன்றவை.

இந்த வடிவில் அறிக்கை Météo நிரலுக்கு அனுப்பப்படும். முடிக்கப்பட்ட மொழிபெயர்ப்பு மூலத்தைப்/அசலைப் போன்று துல்லியமாக வடிவமைக்கப்படும். மேலே கண்ட எடுத்துக்காட்டுக்கு வெளியீடு படம் 3இல் தரப்பட்டுள்ளது.

படம் 3: மெட்டியோ வெளியீடு

U

FPCN11 CYYZ 311630

2

PREVISIONS POUR L ONTARIO EMISES PAR ENVIRONMENT CANADA A 11 H 30 HNE MERCREDI LE MARS 1976 POUR AUJOURD HUI ET JEUDI

1

TORNTO ET BANLIEUE

WINDSOR

0

NAUGEUX AVEC POSSIBIKUTE D AYHIYRD GYU ET JEUDI

0

MINIMUM CE SOIR 4.

-0

HIGH THRUSDAY 10.

2

APERCU POUR VENDREDI…ENSOLELLE.

இந்தக் குறிப்பிட்ட எடுத்துக்காட்டில் மெடியோ உள்ளீட்டின் தவறு காரணமாக *high Thrusday* 10 என்பதை மொழிபெயர்க்கத் தவறியது; HIGHஇல் உள்ள முகப்பெழுத்து/பெரியெழுத்து I என்பதற்குப் பதிலாக எண் 1. தோல்வி தானியக்கமாக ஒழுங்குமுறையால் அடையாளங்காணப்படும் மற்றும் பின் தொகுப்பர்களின்/திருத்தர்களின் (post-editors) கவனத்திற்காக அடையாளப்படுத்தப்படும். ஒழுங்குமுறை தோற்கும் போது தான் மனிதத்தலையீடு ஈடுபடுத்தப்படும்; இது பொதுவாகத் தவறான உள்ளீட்டால் விளைவதாகும். பின் திருத்தியமைப்பு (post-editing) முனையத்தில் குறைந்த அளவு எளிய கட்டளைகளைப் பயன்படுத்திச் செய்யப்படும்.

3.9.3. மொழிபெயர்ப்புச் செயற்பாங்குகள்

Météoவின் மொழியியல் தரவு மரபுச்சொற்கள், இடப்பெயர்கள் மற்றும் பொதுவான (வானிலையியல்) சொற்றொகை என்பவைகளுக்காக மூன்று இருமொழிய அகராதிகளையும் மற்றும் ஆங்கிலத்தின் தொடரியல் பகுப்பாய்வு, பிரஞ்சு மொழியின் தொடரியல் உருவாக்கம்,

பிரஞ்சு மொழியின் உருபனியல் உருவாக்கம் என்பவைகளுக்காக மூன்று செயலாக்கத் தொகுதிகளையும் (processing modules) கொண்டிருக்கும்.

TAUM குழு அதன் முக்கிய இயந்திர மொழிபெயர்ப்புக்கு மாற்றல் அடிப்படையான அணுகுமுறையைப் (transfer-based approach) பின்பற்றினாலும் அடிப்படையில் 'நேரடி' வடிவமைப்பு ('direct' design) Météo துணைமொழி ஒழுங்குமுறைக்கு (sublanguage system) வேண்டி உருவாக்கப்பட்டது. ஆங்கில அறிக்கைகளின் தந்தி நடை அமைப்பு ('telegraphic' style) அடிப்படையில் கிட்டத்தட பிரஞ்சுமொழியின் அறிக்கைகளைப் போன்று இருந்ததன் காரணமாக மாற்றல் தொகுதி (transfer module) தேவையில்லாமல் இருந்தது. மேலும் ஆங்கிலத்தின் உருபனியல் பகுப்பாய்வுக்கு வேண்டிய தொகுதி (module) இல்லாமல் இருந்தது; சொற்றொகையின் பரப்பெல்லை வரையறுக்கப்பட்டிருந்ததுடன் உருபனியல் மாறுபாடுகளும் கட்டுப்படுத்தப்பட்டிருந்தது, எ.கா. வினைகள் நிகழ்கால வடிவிலும் இறந்தகால வினையெச்ச வடிவிலும் மட்டும் தோன்றின.

வானிலை அறிக்கைகளின் கட்டுப்படுத்தப்பட்ட தொடரியல் பரப்பெல்லை காரணமாக மொழியியல் பகுப்பாய்வு ஓரளவுக்கு எளிமைப்படுத்தப்பட்டது: மாற்றுப் பெயர் குறிப்பு, பெயரெச்சத்தொடர், செயப்பாட்டு வாக்கியங்கள் என்பன இல்லை; மேலும் தொடர்கள் குறுகியன. மாறாக, பின்னுருப்புகள், ஆர்டிகிள்/சார்படை இவற்றை விட்டுவிடுவது சிக்கலை உருவாக்கியது; இது ஒழுங்குமுறை உருவாக்கியவர்களை பொருண்மைக் பண்புக்கூறுகள் அடிப்படையிலான பகுப்பாய்வு செயல்முறைகளை இணைத்துக்கொள்ள வழிவகுத்தது.

3.9.3.1. அகராதியில் தேடல் (dictionary look-up)

சொல் தரவின் பிரித்தெடுப்பால் மொழிபெயர்ப்பு தொடங்குகின்றது. எடுத்துக்காட்டக மரபுத்தொடர் அகராதி *blowing snow* என்ற ஆங்கிலத் தொடர் *poudrerie* என்று மொழிபெயர்க்க வேண்டும் என்று சுட்டிக்காட்டும். Météoவைப் பொறுத்தவரையில் ஒரு சொல் போல் எடுத்துக்கொள்ளப்பட வேண்டிய ஒன்றுக்கும் மேற்பட்ட சொற்களின் தொடர்ச்சிகள் யாவும் 'மரபுச்சொற்கள்' ஆகும்; இவ்வாறு 'மரபுச்சொற்கள்' முற்றிலும் நடைமுறை அடிப்படையில் வரையறை விளக்கம் செய்யப்பட்டுள்ளன; கூடுதல் சிறப்பாக இருமொழிய வேற்றுநிலை விதிமுறைகளில் செய்யப்பட்டுள்ளன.

இடப்பெயர் அகராதி ஆங்கிலத்திலும் பிரெஞ்சிலும் வேறுபாடுகள் உள்ள பெயர்கள் (எ.கா. *Newfoundland* → *Terre Neuve*), தொடர்களாக உள்ளவை (எ.கா. *Greater Vancouver* → *Vacouver et banlieue*, *Metro Toronto* → *Toronto et banlieue*) அல்லது சரியான வெளியீட்டிற்காக சில மொழியியல் தகவல்களைக் கட்டாயம் உட்படுத்துவன (எ.கா. பன்மைப் பெயர்கள்,

குறிப்புடைய சார்படை கொண்ட பெயர்கள், பெண்பால் பெயர்கள்) என்பவனவைகளை மட்டுமே கொண்டிருந்தன. நிரல் எல்லா 'அறியப்படாத' சொற்காளையும் இயற்பெயராகக் கொள்ளுவதாலும் எனவே மொழிபெயர்க்காததாலும் பிற எல்லா இடப்பெயர்களையும் விட்டுவிடலாம்.

முக்கிய அகராதி பொது அகராதியாகும்; இது முன்னர் கூறியபடி எல்லா உருபனியல் வடிவங்களையும் கொண்டிருக்கும். ஒவ்வொரு ஆங்கிலச் சொற் பதிவும் பிரஞ்சுமொழி நிகரன், இலக்கண வகைப்பாடு, பொருண்மைப் பண்புக்கூறுகள் மற்றும் இலக்குமொழி உருபனியல் தகவல் என்பனவற்றைக் கொண்டிருக்கும். எடுத்துக்காட்டாக (1),

(1) AMOUNT = N ((F, MSR), QUANTITE)

இதில் N என்பது பெயர், F என்பது பெண்பால் மற்றும் MSR என்பது ஒரு அளவைப் பெயர். இலக்கணவகைப்பாடுகள் மரபு அடிப்படையிலானவை (பெயரடை, வினையடை, இணைப்புக் கிளவி, அடைகொளி அடை போன்றவை).

பெயர்களுக்கு உருபனியல் தகவல் பால் (Fஆ இல்லையா) மற்றும் பன்மை என்பனவற்றை சுட்டிக்காட்டுவது என்ற நிலையில் கட்டுப்படுத்தப்பட்டுள்ளது. பெயரடைகளுக்கு அவை பிரஞ்சுமொழியில் பெயருக்கு முன்னர் வரவேண்டுமா இல்லையா என்பதும் மாற்றங்கள் பெண்பால் மற்றும் பன்மை வடிவுகளை உருவக்க வேண்டுமா என்ற உருபனியல் தகவலும் தரப்பட்டுள்ளன.

வினையடைகளையும் வினைகளையும் பற்றிய தகவல்கள் குறைவானவை ஆகும். வினையடையின் பதிவு அவற்றை பெயரடையுடனா, வினையுடனா, முன்னுருபுகளுடனா சேர்க்கப்படவியலும் என்ற தகவலைக் கொண்டிருக்கும். வினைகள் அவை செயப்படுபொருள் குன்றாவினையா குன்றிய வினையா என்ற தகவலை மட்டும் கொண்டிருக்கும்; வேறு எந்த உள்வகைப்பாட்டுத் (subcategorization) தகவலும் தரப்படவில்லை.

பொருண்மையியல்சார் பண்புக்கூறுகள் பெயர்கள், பெயரடைகள், வினையடைகள், பின்னுருபுகள் என்பனவற்றுடன் இணைக்கப்பட்டிருக்கும். சில எடுத்துக்காட்டுகள் அட்டவணை 1இல் காட்டப்பட்டுள்ளன.

அட்டவணை 1: பண்புக்கூறுகளின் எடுத்துக்காட்டுகள்

பண்புக்கூறு	எடுத்துக்காட்டுகள்
time zone	HAE (heures avacées de l' est)
month	Juillet, novembre
measure	degré, médiocre, supérieur, à peu près, environ
place	secteur, comté, avoisinant, parout, au dessus de
direction	est, du nord
time point	fin, matin, par la suite, avant
time duration	matinée, annuel, pour peu de temps, au cours de
possibility	risque, bon, faible, peut-être
met. phenomenon, stationary	humidité, brume, nuages, chaud, dense
met. phenomenon falling	neig, pluie, grêle, abundant
met. phenomenon blowing	rafle, vent, venteux, fort

மாற்றல் அகராதி (transfer dictionary) இல்லாததால் எல்லா மொழிபெயர்ப்புசார் மாறுபாடுகளும் (translational variants) அகராதியில் பட்டியலிடப்பட்டுள்ளன; அவை பொருத்தமான பெருண்மைப் பண்புக்கூறுகளால் (2) வேறுபடுத்தப்பட்டுள்ளன.

(2) ABOUT = P (MEASURE, ENVIRON)
ABOUT = P (TEMPS, VERB)
AREA = N (LIEU), REGION)
AREA = N (CONDITION METEO), ZONE)
CONSIDERABLE = ADJ ((CONDITON METEO, TOMBANT), FORT)
CONSIDERABLE = ADJ ((CONDITON METEO, STATIONNAIRE), MARQUE

மாறுபாடுகளுக்கு இடையே உள்ள தெரிவு தொடரியல் பகுப்பாய்வின் இறுதி நிலையில் நடைபெறும்.

3.9.3.2. தொடரியல் பகுப்பாய்வு

வானிலை அறிக்கைகளின் ஆய்விலிருந்து, TAUM ஆய்வாளர்கள் அவற்றில் வரும் எல்லாத் தொடர்களின் ஒரு வகைப்பாட்டை நிறுவினர். அவர்கள் ஐந்துவகை கிளையமைப்புகள்

மட்டுமே தேவை என்பதைக் கண்டுகொண்டனர். பகுப்பாய்வு நிரலின் பணி ஒரு தரப்பட்டுள்ள உள்ளீட்டுத் தொடருக்கு ஒரு குறிப்பிட்ட கிளையமைப்பை கண்டுபிடிப்பது என வரையறுக்கப்பட்டுள்ளது.

தொடரியல் பகுப்பாய்வு மூன்று நிலைகளில் (மொழியியல் பார்வை அடிப்படையில்) மரபுத்தன்மையான கீழிருந்து மேல் பகுப்பாய்வு நுட்பதைப் பயன்படுத்தி நிறைவேற்றப்படுகின்றது. மொழியியல் ஆய்வு அடிப்படையில் இதைப் பின்வருமாறு விளக்கவியலும்: இந்த செயன்மை ஒவ்வொரு நிலையிலும் எல்லா சாத்தியமான தீர்வுகளையும் உருவாக்கும்; இறுதி நிலையில் ஒரேயொரு பகுப்பாய்வு வரும்படிக்கு இணைக்கப்படும் போது அவை படிப்படியாகக் குறைக்கப்படும்.

முதல் நிலை நாட்கள், மணிகள், வெப்பநிலை அளபுகள் இவற்றை அறிந்துகொள்வதை உட்படுத்தும். அடுத்த நிலை மிஞ்சும் பெயர்த்தொடர்களை (அதாவது வானிலை நிலைமைகளை வெளிப்படுத்தும் பெயர்த்தொடர்களை) கண்டுகொள்ளும். இந்தச் செயல்முறை சொற்களின் தொடரியல் வகைப்பாட்டையும் அவற்றின் பொருண்மைப் பண்புக்கூறுகளையும் உள்ளடக்கும். இந்தப் பண்புக்கூறுகளை பொருத்துவது சரியான பகுப்பாய்வுக்கும் பிரஞ்சுமொழி மொழிபெயர்ப்புகளுக்கு இடையிலான தெரிவுக்கும் உதவும். இறுதியாக பகுப்பாய்வு வானிலை நிலைமைகள் பிற நிலைமைகளின் நிரப்பிகளாக கலவைத்தன்மையான துணை கிளையமைப்பை உருவாக்கம் என்பதை அறிந்துகொள்ளும். இவ்வாறு தொடரியல் பகுப்பாய்வின் விளைவு ஒரு S கணுவால் ஆதிக்கம் செய்யப்பட்ட தனிக் கிளையமைப்பு ஆகும்; இதில் ஆங்கிலச் சொற்கள் இறுதிக் கணுக்கள் ஆகும்; நிலைபேறாக்கம் செய்யப்பட்ட தொடரமைப்புகளின் வகைகள் இறுதியுறா கணுக்களைக் குறிப்பிடும்.

3.9.3.3.தொடரியல் மற்றும் உருபனியல் உருவாக்கம்

அமைப்புசார் உருப்படுத்தம் (structural representation) மற்றும் சொற்களுடன் இணைக்கப்பட்டுள்ள தகவல்கள் இவற்றிலிருந்து பிரஞ்சுமொழி சொல் வரிசையை ஆக்குவதுதான் தொடரியல் உருவாக்கத்தின் பணி அகும். எடுத்துக்காட்டாக, காலம், இடம் இவற்றின் வெளிப்பாடு வானிலை நிலைமைகளின் வெளிப்பாட்டிற்குப் பின்னர் வரும்; பெயரடைகள் அவை அடைசெய்யும் பெயர்களுக்குப் பின்னர் வரும்; சார்பு அடைகள் செருக்கப்படும். இது பின்னுருபுகளின் சரியான தேவையும் உட்படுத்தும். இறுதி நிலையான உருபனியல் உருவாக்கம் சூழல் சரிகட்டலுடன் பெயரடைகளின் இறுதிகளை (பெண்பால், பன்மை போன்றன) உறுதி செய்வதாகும்.

3.9.4. கணினிசார் செயற்பாங்குகள்

கணினிசார் செயற்பாங்கு அடிப்படையில் Météo ஒரு உற்பத்தி ஒழுங்குமுறையாக ஒரு தனி ஒருங்கிணைக்கப்பட்ட ஒழுங்குமுறையாக அமைக்கப்பட்டுள்ளது. எல்லாத் தொகுதிகளுக்கும் ஒரு தனியான தரவு அமைப்பும் விதி எழுதும் வடிவவாதவுமே உள்ளன. தரவு அமைப்பு சார்ட் (chart) வடிவில் உள்ளது; அவற்றின் விற்கள் பண்புக்கூறுகளின் கட்டுக்களால் புலக்குறிப்பு

செய்யப்பட்டுள்ளன. இலக்கண விதிகளின் பிரயோகத்தின் வரிசைமுறையைத் தீர்மானிக்க வெளிப்படையான கட்டுப்பாடு இல்லை. பாரம்பரிய உற்பத்தி ஒழுங்குமுறைகளின் வழக்கமான முறையில் அவை விளக்கப்பட்டுள்ளன. உற்பத்தி ஒழுங்குமுறை நிர்மாணம் கீழிருந்து-மேல் அகலம்-முதல் பகுப்பாய்வியுடன் ஒத்திருக்கும். மென்பொருள் Alain Colmerauer என்பவரால் எழுத்தப்பட்டுள்ளது; இது Q-ஒழுங்குமுறை (Q என்பது Quebec என்பதைக் குறிப்பிடும்) என்று அழைக்கப்படுகின்றது. Q-ஒழுங்குமுறை இதையொத்த கணினிசார் நிர்மாணத்தைக் கொண்டுள்ள ப்ரோலாக் (Prolog) என்ற நிரலாக்க மொழியின் மூல முன்வகையாகக் (prototype) கருதப்படுகின்றது.

3.9.4.1. தரவு அமைப்பு

Météoவின் தரவு அமைப்பு ஒரு சார்ட் ஆகும்; அதன் விற்கள் பண்புக்கூறுகளின் கட்டுக்களால் புலக்குறிப்பு செய்யப்பட்டிருக்கும். அகராதி தேடல்/பார்த்தல் ஒவ்வொரு சொற்களுகளின் அர்த்தத்திற்கும் ஒரு வில்லை உருவாக்கி சார்ட்டின் தொடக்கக் கட்டமைப்பை (intial configuration) உருவாகும்.

கணினி அடிப்படையில் பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் இடையில் வேறுபாடு இல்லை. பகுப்பாய்வைப் போன்று உருவாக்கம், சார்டின் மேல் அமைவுப் பொருத்தத்தை (pattern match) உள்ளடக்கிய விதிகளாலும் கூடுதல் விற்களின் கட்டுமானத்தாலும்/உருவாக்கத்தாலும் உயிர்ப்பிக்கப்படுகின்றது. இருப்பினும் இந்த இரு செயற்பாங்குகளும் தனியாக வைக்கப்பட்டுள்ளன. பாகுபாட்டின் இறுதியில் சார்ட் ஒழுங்குபடுத்தப்பட்டு சார்ட் அடுத்த உருவாக்க தொகுதிக்கு அனுப்பபடுவதற்கு முன் தேவையில்லாத விற்கள் நீக்கப்படும். இது கணினி திறனை அதிகரிப்பதற்கான உபாயமேயன்றி ஒட்டுமொத்த மொழிபெயர்ப்பு செயல்முறையின் கண்டிப்பான தொகுதியாக்கம் அல்ல.

3.9.4.2. **விதி வடிவவாதம்**

விதிகள் GETA விதிகளைப் போன்றன; அவை ஒரு கிளையமைபை விளக்கும்; பின்னர் கிளையமைப்பின் ஒவ்வொரு கிளையிலும் பண்புக்கூறுகளை விளக்கும். விதியின் இடப்பக்கம் சார்ட்டில் உள்ள விற்களின் வரிசை ஆகும்; வலப்பக்கம் புதிய வில்லை குறிப்பிடும். எடுத்துக்காட்டாக பின்வரும் விதி worse என்பதை பகுதி கூட்டல் ஒரு பின்னொட்டு –er என்ற உருபனியல் விதியைப் பொருள்கோள் செய்கின்றது.

WORSE == ADJ (BAD, /) + * (ER)

விதி இடது பக்கத்தில் WORSE என்ற கிளையமைப்பை வரையறை விளக்கம் செய்கின்றது மற்றும் வலது பக்கத்தில் இரு கிளைகளைக்கொண்ட ஒரு கிளையமைப்பை உருவாக்குகின்றது;

ஒன்று துணைநிலை தனிமம் BAD என்பதைக் கொண்ட ADJ மற்றும் துணைநிலை ER கொண்ட *.

3.9.4.3. **விதிப் பிரயோகம்**

Q-ஒழுங்குமுறை ஒரு பாரம்பரிய உற்பத்தி ஒழுங்குமுறையாகும்; விதித் கணம்/குழுமம் (rule-set) இந்த மொழியியல் விதிகளின் வரிசைப்படுத்தப்படாத கணமாகும். இந்த செயற்பாங்கு தரவுத்தளத்தின் (சார்ட்) நிலையால் கட்டுபடுத்தப்பட்டுள்ளது. இது விதித்தொகுப்பில் உள்ள எந்த விதியை அடுத்தப்படியாகப் பயன்படுத்த இயலும் என்பதைத் தீர்மானிக்கும். பெரும்பாலும் ஒன்றிற்கு மேற்பட்ட விதிகளை பயன்படுத்த இயலும்; அப்பொழுது இரண்டு சாத்தியங்களும் கோட்பாடு அடிப்படையில் இணையாகப் பயன்படுத்தப்பட வேண்டும்.

மொழியியல் செயலாக்கத்தின் இந்தக் கட்டமைப்பற்ற அணுகுமுறை முக்கியமான குறைப்பாடுகளைக் கொண்டுள்ளது; இவை Météoவில் அதன் இறுக்கம் காரணமாகத் தவிர்க்கப்பட்டுளன. ஒப்பீட்டளவில் விதித்தொகுதி சிறியதாகும்; இதன் காரணம் பனுவல் வகை மிகவும் கட்டுப்படுத்தப்பட்டது. Météoவின் உற்பத்தி ஒழுங்குமுறை நிர்மாணம் பொருத்தமாக இருந்தாலும் இது மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்குப் பொதுவாகப் பொருந்தக்கூடியனவற்றைப் பின்பற்றவில்லை. TAUM ஆய்வாளர்கள் இதை அவர்களின் மிக விரிவான இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறைக்குப் பயன்படுத்த முயன்றபோது கண்டுகொண்டார்கள்.

ஒரு தனி ஒருங்கிணைக்கப்பட்ட விதி எழுதும் வடிவவாத்தையும் அதனுடன் தொடர்புடைய கணினி நுட்பத்தையும் பயன்படுத்துவதில் பல சாதியமான குறைபாடுகள் உள்ளன. அகராதி தேடல், பகுப்பாய்வு, உருபனியல் உருவக்கம் என்பன யாவும் ஒரே விரிதெழுது விதி உற்பத்தி ஒழுங்குமுறை உத்தியால் செய்யப்படுகின்றன. இந்த நுட்பம் இப்பணிகளில் மிகக் கலவைத்தன்மையான ஒன்றான பகுப்பாய்விற்கு இடமளிக்கப் போதுமான திறனைக் கொண்டிருக்கவில்லை. ஆனால் பொதுவான விரித்தெழுது நுட்பத்தின் எல்லாத் திறனையும் எளிய பணியான அகராதி தேடலுக்கும் உருபனியல் உருவாகத்திற்கும் பயன்படுத்துவது அதிகப்படியானதாகத் தோன்றும். Météoவின் குறைந்த அளவு காரணமாக இது அவ்வளவு பெரிய விஷயமல்ல; ஆனால் பெரிய ஒழுங்குமுறைகளுக்கு Météoவின் இந்த அம்சங்கள் விரும்பக்கூடியன அல்ல.

3.9.5. முடிவுரை

Météo தெளிவாக இரண்டாவது தலைமுறை மொழிபெயர்ப்பு ஒழுங்குமுறை என்றாலும் சுவாரஸ்யமான விதங்களில் இது பாரம்பரிய நிர்மாணியத்திலிருந்து விலகிச் செல்கின்றது. இதன்

மொழியியல் உபாயங்கள் நல் ஊக்கப்படுத்தப்படவை மற்றும் கையிலுள்ள பணிக்கு வேண்டி சீரமைக்கப்பட்டுள்ளது. கணினி அடிப்படையில் இது அதிநவீனமானது மற்றும் அதன் காலத்திற்குப் புதுமையானது. அடிப்படை வடிவாக்கத்தில் Météo ஒரு நேரடி ஒழுங்குமுறையாகும்.

3.10. வெயிண்டர் தகவல்தொடர்பு நிறுவனம் WEIDNER COMMUNICATIONS CORPORATION

வெயிண்டர் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை புரூஸ் வெயிண்டரால் 1977-இல் நிறுவப்பட்டது (Slocum, 1985: 8). வெயிண்டர் சில இயந்திர மொழிபெயர்ப்பு நிரலர்களை (programmers) ப்ரிகாம் யங் பல்கலைக்கழகத்திலிருந்து வேலைக்கு அமர்த்தினார். அவர் ஆங்கிலம்-பிரஞ்சு ஒழுங்குமுறையை உருவாக்கி 1980-இல் கனடாவிலுள்ள Mitel-உக்குக் கொடுத்தார். அதே ஆண்டு பீட்டா-டெஸ்ட் ஆங்கில-ஸ்பானிஷ் ஒழுகுமுறையைச் சிமென்ஸ் கார்பரேஷனுக்கு வழங்கினார். 1981-இல் மிடெல் வெயிண்டரின் ஆங்கில-ஸ்பானிஷ், ஆங்கில-ஜெர்மன் ஒழுங்குமுறைகளை வாங்கியது. ப்ரேவைஸ் (ஜப்பானிலுள்ள ஒரு மொழிபெயர்ப்பு சேவைச் செயலகம்) வெயிண்டரின் ஆங்கில-ஸ்பானிஷ், ஆங்கில-ஜெர்மன் ஒழுங்குமுறைகளை வாங்கியது. வெயிண்ட மொழிபெயர்ப்பு ஒழுங்குமுறை "முழு தானியங்கு ஒழுங்குமுறையாக" இருந்தபோதிலும் அது மொழிபெயர்ப்புக்கு "இயந்திர உதவி" என்று வணிகம் செய்யப்பட்டது. இது (பனுவல்களின் சொல்சார் முன்பகுப்பாயு, அகராதிகளின் உருவாக்கம், போன்ற) பிற நோக்கங்களுக்கு மிக அதிக அளவிலான ஊடாட்டச் செயல்பாடு கொண்டுள்ளது மற்றும் இது சொல்லாய்வு மென்பொருளை வெளிக் கருவிகளுடன் (எடுத்துக்காட்டாக மிட்டலில் ஜெராக்ஸ் 9700 லேசர் அச்சுப்பொறி) ஒருங்கிணைக்கின்றது. இவ்வாறு வெயிண்டர் மொழிபெயர்ப்பு ஒழுங்குமுறை வடிவாக்கம் செய்யப்பட்ட மூலமொழி ஆவணங்களை ஏற்றுக்கொண்டு வடிவாக்கம் செய்யப்பட்ட மொழிபெயர்ப்பை உருவாக்குகின்றது. எல்லோரும் வடிவாக்கம் செய்யப்பட்ட மூலப் பனுவல்களிலிருந்து வடிவாக்கம் செய்யப்பட்ட மொழிபெயர்ப்புகளை உருவாக்க விரும்புவதால் இது பயனாளிகளுக்கு முக்கியமான சிறப்பம்சமாகும்.

இந்த மொழிபெயர்ப்பு ஒழுங்குமுறை நவீனச் சொல் செயற்பாங்குத் தொழில்நுட்பத்துடன் (modern word-processing technology) இறுக்கமாக ஒருங்கிணைக்கப்பட்டுள்ளது என்பதன் காரணமாக மொழிபெயர்ப்புக் கூறு எந்த அளவுக்கு மொழிபெயர்ப்பு உற்பத்தியை அதிகரிக்கிறது அல்லது முந்தைய மனித முயற்சியின் (அல்லது மோசமாக தானியக்கம்செய்யப்பட்ட) செயற்பாங்குகளின் எளிய தானியக்கம் எந்த அளவுக்கு உற்பத்தி பேறுக்கு உதவுகின்றது என்பதை மதிப்பீடுசெய்வது கடினமாகின்றது. நேரடியான

மொழிபெயர்ப்புக் கூறு குறிப்பாக அதிநவீனமானதல்ல. குறிப்பாக, பகுப்பாய்வு குறிப்பிட்ட எல்லைக்குட்பட்டதாகும்; பெரும்பாலும் பெயர்த்தொடர் அல்லது வினத்தொடர் நிலைகளுக்குக் கட்டுப்படுத்தப்பட்டதாகும். உயர்நிலைகளில் இருக்கிற சூழலைக் கணக்கில் எடுக்க இயலாது.

மொழிபெயர்ப்பு நான்கு சுதந்திரமான நிலைகளில் செய்யப்படுகிறது: ஒப்பெழுத்து பொருண்மைமயக்க நீக்கம், மரபுத்தொடர் தேடல், கட்டமைப்பு பகுப்பாய்வு, மாற்றல். இந்நிலைகள் கூடுதல் சிக்கல்கள் உருவாகும் என்ற காரணத்தால் ஒன்றுக்கொன்று ஊடாடுவதில்லை.

1982-இல் வெயிண்டர் கிரேட் பிரிட்டனில் உள்ள ITT-க்கு ஆங்கில-ஜெர்மன், ஜெர்மன் - ஆங்கில இயந்திர ஒழுங்குமுறைகளை வழங்கியது. வெயிண்டர் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை வியாபாரம் செய்துகொண்டிருந்தது. வியாபார ரீதியான ஜப்பானீஸ்-ஆங்கில ஒழுங்குமுறையை உருவாக்க முயன்றது. ஒரு வியாபார ரீதியான ஜப்பானீஸ்-ஆங்கில மொழிபெயர்ப்பு ஒழுங்குமுறை ப்ராவைசால் (Bravice) அறிவிக்கப்பட்டு தொடங்கப்பட்டது. குறிப்பாக, பிடிபி-11 (PDP-11) மீதும் பொதுவாக ஏதோ ஒரு கணிப்பொறி மீதும் அதன் சார்பைக் குறைக்க வெயிண்டர் தனிநபர் கணினி ஐபிஎம் மீது அதன் மொழிபெயர்ப்பு ஒழுங்குமுறையை நடைமுறைப்படுத்தியது. இந்த ஒழுங்குமுறை சில அசம்பிளி குறிய ஆதரவுடன் ஃபோர்ட்ரானில் எழுத்தப்பட்டது. இதன் நெகிழ்வுத்தன்மையை அதிகரிக்க வேறு மொழியில் இந்த மென் பொருளை எழுதும் திட்டமும் நடைமுறையில் உள்ளது.

3.11.ஸ்பனாம் (SPANAM)

ஒரு நம்பிக்கைக்குரிய செயலாக்க ஆய்வைத் தொடர்ந்து, வாஷிங்டனில் டிசியில் (Washington, D.C.) உள்ள பான் அமெரிக்கன் சுகாதார அமைப்பு (Pan American Health Organization), 1975இல் GAT உருவாக்கப்பட்ட அதே உத்திகள் சிலவற்றைப் பயன்படுத்தி ஒரு இயந்திர மொழிபெயர்ப்பு உருவாக்க வேலையைத் தொடங்க முடிவுசெய்தது (Slocum, 1985: 9).

அருகிலுள்ள GAT-இன் உறைவிடமான ஜார்ஜ்டவுன் பல்கலைக்கழகத்தில் இருந்து ஆலோசகர்கள் அமர்த்தப்பட்டனர் அதிகார பூர்வமான PAHO மொழிகள் ஆங்கிலம், பிரஞ்சு, போர்த்துகீசியம், ஸ்பானிஷ் ஆகியனவாகும்; ஸ்பானிஷ்-ஆங்கிலம் தொடக்க மொழி இணையாக பின்வரும் நம்பிக்கை காரணமாகத் தேர்வு செய்யப்பட்டன: "சமாளிக்கவியலும் வெளியீட்டை உருவாக்க இந்த சேர்க்கை குறைவான பாகுபடுத்த உத்திகளையே வேண்டும் (மற்றும் மொழியியல் விதிகள் விட மென்பொருளுக்கு செலவிடும் முயற்சி தொடர்பான காரணங்கள்)"(Vasconcellos 1983). உண்மையான பணி 1976 ஆம் ஆண்டு தொடங்கப்பட்டது, முதல் முன்மாதிரி 1979இல் ஐ.பி.எம். மெயின்பிரேமில் பஞ்கார்ட் உள்ளீட்டைப் பயன்படுத்தி செயல்பட்டது. இதைத் தொடர்ந்து ஒரு சொல்-செயற்பாங்கு ஒழுங்குமுறையின் (word-processing

system) ஒருங்கிணைப்பு இயந்திர மொழிபெயர்ப்பை உருவாகும் பயன்பாட்டிற்கு முக்கியமாகக் கருதப்பட்டது.

கூடுதல் மேம்படுத்தலுக்குப் பின்னர் SPANAM அடிப்படையில் நிறுவனத்திற்குள்ளான மொழிபெயர்ப்புச் சேவை 1980-ஆம் ஆண்டு உருவாக்கப்பட்டது. 1980ஆம் ஆண்டு முதல், SPANAM பனுவலின் ஒரு மில்லியன் வார்த்தைகளுக்கு மேல், சராசரியாக நாளைக்கு சுமார் 4,000 வார்த்தைகளை மொழிபெயர்க்கப் பயன்படுத்தப்பட்டது.

பிந்தைய திருத்தாளர்கள் (post-editors) திருத்தப்பணி வேகமாக நடைபெற பல தந்திரங்களைக் கைவசம் வைத்திருந்தனர்; SPANAM ஆங்கில வெளியீட்டைக் கையாள சொல் செயலியில் (word processor) சிறப்புக் கோர்வை செயல்பாடுகள் (special string functions) உட்படுத்தப்பட்டன.

SPANAM ஆரம்ப நிலை குறித்து மேலோட்டமான விவரங்கள்: அடித்தளத்தில் இருக்கும் மொழியியல் தொழில்நுட்பம் GAT அடிப்படையில் இருந்தது; இலக்கண விதிகள் நிரல்களில் GAT பாரம்பரியத்தில் உருவாக்கப்படது. நிரல்கள் கட்டகமாக இருக்கும்படி மென்பொருள் தொழில்நுட்பம் புதுப்பிக்கப்பட்டது. மொழிபெயர்ப்பு ஒழுங்குமுறை அதிநவீனமானதல்ல: இது நேரடி மொழிபெயர்ப்பு உத்தியைக் கையாண்டது; பழமையான சுதந்திரமான செயலாக்க நிலைகளின் வரிசை வழியாக தொடர்கள் மற்றும் எச்சத்தொடர்களின் எல்லைக்குட்பட்ட பகுப்பாய்வு செய்யப்பட்டது. SPANAM தற்போது மூன்று PAHO மொழிபெயர்ப்பாளர்களால் தங்கள் வழக்கமான வேலையைச் செய்யப் பயன்படுத்தப்படுகிறது

ENGSPAN (ஆங்கிலம்-ஸ்பானிஷ்) உருவாக்க ஒரு தொடர் திட்டம் 1981ஆம் ஆண்டில் இருந்து தொடங்கப்பட்டு, இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை உற்பத்திக்கு வழங்கப்பட்டுள்ளது; கூடுதல் மேம்பட்ட வடிவமைப்பு (எ.கா. ஏடிஎன் பாகுபடுத்தி (ATN parser)) இதன் சிறப்பம்சம் ஆகும்; சில அம்சங்கள் SPANAMஇல் உட்படுத்தப்பட்டுள்ளன. இந்த இரண்டு மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் வெற்றி காரணமாக (போர்த்துகீசியத்தை இலக்கு மொழியாக கொண்டு) ENGPORT என்ற ஒழுங்குமுறை தொடங்கப்பட்டு நடைமுறைப் படுத்தப்பட்டுள்ளது.

3.12. கல்ட்: சீன பல்கலைக்கழக மொழி மொழிபெயர்ப்பாளர் (CULT: CHINESE UNIVERSITY LANGUAGE TRANSLATOR)

இயந்திர உதவி பெறும் மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் மிக வெற்றிகரமான ஒன்றாக CULT-ஐ கருத இயலும். இதன் உருவாக்கம் 1968இல் ஹாங்காங் சீன பல்கலைக்கழகத்தில் தொடங்கப்பட்டது (Slocum, 1985: 9). CULT (பெய்ஜிங்கில் வெளியிடப்பட்ட) சீனக் கணிதம் மற்றும் இயற்பியல் இதழ்களை ஒரு உயர்ந்த ஊடாடும் செயற்பாங்கு மூலம் (அல்லது, குறைந்த பட்சம், அதிகமான மனித தலையீடு மூலம்) ஆங்கிலத்தில்

மொழிபெயர்க்கும். இதன் இலக்கு உள்ளீட்டில் பெரிய அளவு முன் திருத்தத்தையும் மொழிபெயர்ப்பின் போது (உள்ளீட்டில் வாக்கியங்களின் மற்றும் தொடர்களின் எல்லைகளை அடையாளப்படுத்துவது, தேவைப்படும் போதெல்லாம் சொற்பொருளை குறிப்பிடுவது போன்ற) ஒரு குறிப்பிட்ட அளவு மனிதத் தலையீட்டை அனுமதிப்பதன் மூலம் வெளியீட்டில் பின் திருத்தத்தை அகற்றுவதாகும். CULT ஐசிஎல் 1904A கணினியில் இயங்கியது.

1975ஆம் ஆண்டு தொடக்கத்தில் CULT ஆக்டா மேதமெடிகா சிநிக்காவை (*Acta Mathematica Sinica*) ஆங்கிலத்தில் மொழிபெயர்க்கும் பணிக்குப் பயன்படுத்தப்பட்டது; 1976ஆம் ஆண்டு ஆக்டா ஃபிசிக்கா சிநிக்காவையும் (Acta Physica Sinica) மொழிபெயர்ப்பது இணைக்கப்பட்டது. ஒரு நூற்றாண்டுக்கு முன்பு கண்டுபிடிக்கப்பட்ட தந்தி குறிங்களைப் பயன்படுத்தி முதலில் சீன எழுத்துக்களை வரிவடிவாக்கம் செய்யும் சிக்கல் தீர்க்கப்பட்டது; உள்ளீட்டுத் தரவு அட்டைகளில் குத்தப்பட்டது. ஆனால் 1978ஆம் ஆண்டு ஆன்-லைனில் தரவை உள்ளீடுசெய்வதற்கும் முன் மற்றும் பின் திருத்தத்திற்கும் வேண்டி சொல் செயற்பாங்கு கருவிகள் கூடுதலாகச் சேர்க்கப்பட்டு மொழிபெயர்ப்பு ஒழுங்குமுறை மேம்படுத்தப்பட்டது.

CULT-இன் பின்னணியிலுள்ள நுட்பங்கள் எந்த அளவுவுக்குப் பொதுவானவை என்பது தெளிவாகத் தெரியவில்லை; எடுத்துக்காட்டாக, அதைப் பிற நூல்களை மொழிபெயர்க்கப் பயன்படுத்த இயலுமா, இதன் செயல்பாடு சிக்கனமானதா போன்றவை தெளிவாகத் தெரியவில்லை (Slocum, 1985: 10).

3.13.ஆல்ப்ஸ்: தானியங்கி மொழிச் செயற்பாங்கு ஒழுங்குமுறைகள் (ALPS: AUTOMATED LANGUAGE PROCESSING SYSTEMS)

ஆல்ப்ஸ் 1980இல் பிரிக்ஹாம் யங் பல்கலைகழக (Bringham Young University) ஐந்து ஐடிஎஸ் மேம்பாட்டாளர்கள் (ITS developers) அடங்கிய ஒரு குழுவால் ஒன்றிணைக்கப்பட்டது. இந்தக் குழு மனித மொழிபெயர்ப்பாளர்களுக்கு (அகராதி பார்த்தல் மற்றும் பதிலீட்டு செய்தல் முதலிய) இயந்திர உதவிக் கருவிகளை உற்பத்திசெய்வதில் ஆர்வம் உள்ள மொழியியலாளர்களைக் கொண்டிருந்தது. பின்னர் ஐடிஎஸ் ஊழியர்களிலிருந்து முக்கிய நபர்கள் அனைவரும் சேர்ந்துகொண்டனர். இவ்வாறு புதிய ஆல்ப்ஸ் ஒழுங்குமுறை அனைத்து விதத்திலும் ஊடாடும் கருவியாக அமைந்தது; மற்றும் தீவிரமாக மொழிபெயர்ப்பு செய்வதாகப் பாசாங்கு செய்யவில்லை; மாறாக, ஆல்ப்ஸ் அன்றாட மொழிபெயர்ப்பு அனுபவத்தில் எதிர்கொள்ளப்படும் பணிகளை தானியக்கம் செய்ய ஒரு தொகுதி மென்பொருள் கருவிகளை மொழிபெயர்ப்பாளர் வழங்குகியது (Slocum, 1985).

ஆல்ப்ஸ் BYU ITS மொழிபெயர்ப்பு ஒழுங்குமுறைகள் ஆதரவு கொடுத்த மொழி இணைகளை (ஆங்கிலத்திலிருந்து பிரஞ்சு, ஜெர்மன், போர்த்துகீசியம், மற்றும் ஸ்பானிஷ் மொழிபெயர்ப்பதை) சுவீகரித்துக்கொண்டது. பின்னர், பிற மொழிகள் (எ.கா., அரபு)

அறிவிக்கப்பட்டுள்ளன; ஆனால் அவற்றின் வணிக நிலை தெளிவாகத் தெரியவில்லை. இயந்திர உதவி மொழ்பெயப்பு ஒழுங்குமுறைகளை விற்பத்துடன் கூடுதலாக, ஆல்ப்ஸ் தற்போது தனது சொந்த மொழிபெயர்ப்புச் சேவைப் பிரிவை உள்ளடக்கும் (Slocum, 1985).

புதிய ஆல்ப்ஸ் மொழிபெயர்ப்பு ஒழுங்குமுறை எந்த மூன்று "நிலைகளிலும்" வேலை செய்யும் எனக் கருதப்பட்டது - பன்மொழி சொல் செயற்பாங்கிலிருந்து திறன்களை வழங்குவது மற்றும் அகராதியில் நோக்குவது, சொல்லுக்குச் சொல் மொழிபெயர்ப்பிலிருந்து மிகவும் தானியங்கித் தன்மையான (மனித உதவி வழி) வாக்கிய-நிலை மொழிபெயர்ப்பு வரை. ஆல்ப்சால் வழங்கப்பட்ட மையக் கருவி ஆன்-லைன் அகராதியால் இணைக்கப்பட்ட ஒரு பட்டியலால் உந்தப்படும் சொற்செயற்பாங்கு ஒழுங்குமுறை ஆகும். ஆல்ப்ஸ் மொழிபெயர்ப்பு ஒழுங்குமுறையும் பல நிறுவங்களால் பயன்படுத்தப்பட்டது.

3.14. ஜப்பானில் இயந்திர மொழிபெயர்ப்பு வளர்ச்சி

1982ஆம் ஆண்டில் ஜப்பான் அதன் புதிய ஐந்தாம் தலைமுறை ஆய்வுத்திட்டத்தைப் பிரசுரித்தும் புதிய தலைமுறை கணிப்பொறி தொழில்நுட்ப (New Generation Computer Technology (ICOT)) நிறுவனம் நிறுவியும் தொழில்நுட்ப உலகை திடுக்கிடச்செய்தது. அதன் நோக்கம், மேற்கத்திய தொழில்நுட்பத்தைத் ஒரேயடியாகத் தாண்டி 1990களில் டிஜிட்டல் மின்னணு உலகில் முன்னணியில் ஜப்பானை இருத்துவதாகும் (Slocum, 1985).

ஜப்பான் நாட்டின் சர்வதேச வர்த்தக மற்றும் தொழில் துறை அமைச்சுதான் (Japan's Ministry of International Trade and Industry (MITI)) இத்திட்டத்தை ஊக்குவித்த சக்தி ஆகும்; இது கணிப்பொறி கட்டமைப்பு மற்றும் செயற்கை நுண்ணறிவில் மிகவும் புதுமையான நுட்பங்களின் வளர்ச்சி மற்றும் பயன்பாட்டின் வழி நோக்கத்தை அடைய வேண்டும் என்று விரும்பியது.

ICOT விஞ்ஞானிகள் மற்றும் பொறியாளர்களால் பயன்பாட்டுப் பகுதிகளில் ஒன்றாகக் கருதப்பட்ட இயந்திர மொழிபெயர்ப்பு (Moto-oka 1982) ஒரு முக்கிய பங்கை வகிக்கின்றது. மேற்கத்திய செயற்கை அறிவு அறிவியலார்கள் இடையில் இயந்திர மொழிபெயர்ப்பைச் சேர்ப்பது பொருத்தமில்லாததாகக் கருத்தப்பட்டதாகத் தோன்றுகின்றது: ALPAC நெருக்கடிக்குப் பின் இயந்திர மொழிபெயர்ப்பில் வெற்றி கிடைக்காது என்று உலக அளவில் நம்பப்பட, இரண்டு பதின்ம ஆண்டுகள் செயற்கை அறிவு அறிவியலார்கள் இயந்திர மொழிபெயர்ப்பைப் புறக்கணித்தார்கள். அனால் ஜப்பான் தலைமை இயந்திர மொழிபெயர்பை உட்படுத்தியது விபத்தல்ல. ஜப்பானிய மொழிக்கு மொழிபெயர்ப்பது ஜப்பானிய ஆய்வாளர்களுக்கு அவர்களை ஒத்த மேற்கட்த்திய ஆய்வாளர்களிடமிருந்து தகவல்களைப் பெறுவதற்கு முக்கியமான வழியாக அமைந்தது. ஜப்பானியர்கள் மொழிபெயர்ப்பை அவர்களுடைய தொழிநுட்ப நிலைநிற்புக்கு தேவையானதாகக் கருதினர்; ஆனால் அதை மிகக் கடினமான ஒன்றாக உணர்ந்தனர். மிகப் பெரிய மூலதனம் மொழிபெயர்ப்புக்காக அவர்களால் செலவிடப்பட்டது.

18க்கும் கூடுதலான குழுக்களுகள் இயந்திர மொழிபெயர்ப்பில் ஈடுபட்டனர் (Nomura 1982). ஜப்பானிய ஆய்வுத்திட்டங்களில் பல மிகப்பெரியதாக அமைந்தது. பல ஜப்பானிய ஆய்வுத்திட்டங்கள் மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதுடன் ஆய்விலும் ஈடுபட்டிருந்தனர். இயந்திர மொழிபெயப்பில் ஜப்பானியரின் முன்னேற்றம் உடனடியாக விளைய வில்லை. அவர்களின் வன்பொருளின் தரக்குறைவும் மென்பொருளின் திறமை குறைவும் இதற்குக் காரணமாக அமைந்தன. மிகுந்த கடினமான உழைப்புக்குப் பின்னர் முன்னேற்றம் அடைந்தனர். சில பயன்பாடுகளின் உற்பத்தி தொடங்கப்பட்டது. ஜப்பானிய மொழிபெயர்ப்பு ஒழுங்குமுறைகள் எதுவும் நேரடியானதல்ல (Slocum, 1985); எல்லாம் உலகமயமான ஆய்வில் ஈடுபட்டன; பெருபாலானவை மாற்றல் அணுகுமுறையாக (transfer approach) அடிப்படையாகக் கொண்டு அமைந்தன; சில குழுக்கள் இடைமொழி அணுகுமுறையில் (interlingua approach) ஈடுபட்டிருந்தன.

1964-இலிருந்து இயந்திர மொழிபெயர்பு ஆய்வு கொயொட்டொ பல்கலைக்கழகத்தால் (Koyoto Univeristy). ஒரு காலகட்டத்தில் இரண்டு இயந்திர மொழிபெயர்ப்பு ஆய்வுத்திட்டங்கள் நடைமுறையில் இருந்தன. அண்மையில் கைவிடப்பட்ட ஆங்கில-ஜப்பானிய மொழி ஆய்வுத்திட்டம் முறைசார்ந்த பொருண்மையியல் அடிப்படையில் அமைந்தது. பிந்தையது அறிவியல் மற்றும் தொழில்நுட்பக் கட்டுரைகளின் ஆங்கிலத் தலைப்புகளை மொழிபெயர்க்க ஒரு நடைமுறை ஒழுங்குமுறையாக உருவாக்கப்பட்டது.

இருப்பினும் பெரும்பாலான ஜப்பானிய இயந்திர மொழிபெயர்ப்புகள் தொழிற்சாலை ஆய்வுக்கூடங்களில் செய்யப்பட்டன. புஜித்சூ (Fujitsu) (Swai et al), ஹிட்டாச்சி (Hitachi, தோஷிபா (Toshiba) (Amano 1982) மற்றும் நெக் NEC (Muraki & Ichiyamana 1982) என்பன பொதுவாக கணினி கையேடுகளை மொழிபெயர்க்க பெரிய திட்டங்களுக்கு ஆதரவு அளித்தன. நிப்பான் டெலிகிராப் மற்றும் தொலைபேசி (Nippon Telegraph and Telephone), அறிவியல் மற்றும் தொழில்நுட்ப கட்டுரைகளை ஜப்பனிய மொழியிலிருந்து ஆங்கிலத்தில் மொழிபெயர்க்க ஒரு ஒழுங்குமுறையை உருவாக்க முயன்றனர் (Nomura et al 1982). எதிர்காலத்தில் தொலைபேசி உரையாடலை உடனடியாக மொழிபெயர்ப்பதை எதிர்காலத் திட்டமாகக் கொண்டிருந்தது.

ஜப்பான், இயந்திர இயந்திரமொழிபெயர்ப்பு தொழிநுட்பத்தைப் பெறவும் உருவாக்கவும் நீண்டகாலப் பொறுப்பான செயல்பாட்டை கொண்டுள்ளது.

3.15. ஏரியன் (கேதா) (Ariane (GETA))

இந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை Arine-78, Arine-85, Arine-G5 என பல முன்னேற்றங்களையும் மாற்றங்களையும் அடைந்துள்ளது. இதன் நீண்டகால இலக்கும் ஒரு

பன்மொழிய மொழிபெயர்ப்புக்கு இயந்திரமொழிபெயர்ப்பு இயந்திரம் (MT 'engine') என்பதை உருவாக்குவதாகும் (Hutchins and Somers, 1992:222).

3.15.1. **வரலாற்றுப் பின்னணி**

முன்னர் கூறியபடி கிரெனோபிள் CETA ஒழுங்குமுறையை கைவிடப்பட்ட போது Groupe d'Etudes pour la Traduction Automatique (GETA) என்பது உருவானது (Slocum, 1985). இடைமொழி அணுகுமுறையின் தோல்வியின் எதிர்வினையாக GETA மாற்றல் அணுகுமுறையை சுவீகரித்தது. மேலும் முந்தைய மென்பொருள் வடிவமைப்பின் பெரும்பான்மையும் கைவிடப்பட்டது; செயற்பாங்கின் ஒரு புதிய பாணியை ஆதரிக்கும் புதிய மென்பொருள் தொகுப்பால் இடம்பெயர்க்கப்பட்டது. GETA மொழிபெயர்ப்பு ஒழுங்குமுறையின் முக்கியப் பகுதி (ஏரியன்-78) மூன்று வகையான நிரல்களால் உருவாக்கப்பட்டது.

ஒன்று (எடுத்துக்காட்டாகச் சொல் பகுப்பாய்வுக்கு வேண்டி) கோர்வைகளைக் கிளைகளாக மாற்றும், இரண்டாவது (எடுத்துக்காட்டாக தொடரியல் பகுப்பாய்வு மற்றும் மாற்றலுக்கு வேண்டி) கிளையமைப்பை கிளையமைப்புகளாக மாற்றும் மூன்றாவது (எடுத்துக்காட்டாக சொல் உருவாக்கத்திற்கு வேண்டி) கிளைகளைக் கோர்வைகளாக மாற்றும். ஒட்டுமொத்தமான மொழிபெயர்ப்புச் செயற்பாங்கும் பல நிலைகளின் வரிசையால் உருவாக்கப்பட்டது; இதில் ஒவ்வொரு நிலையும் இந்நிரல்களில் ஒன்றைப் பயன்படுத்தும். ஏரியன்-78-இல் உள்ள பிற தொகுதிகள் திருத்துவதையும் ஒழுங்குமுறையை பராமரிக்கும் செயல்பாட்டையும் செய்யும்.

பிற இயந்திர மொழிபெயர்ப்புத் திட்டங்களிலிருந்து ஏரியன்-78 வேறுபடுத்தும் அம்சங்களில் ஒன்று எந்த நிலையும் அது சரியாக செயல்பட தேவையான குறைந்த அளவு திறனிலிருந்து அதிக திறன்வாய்ந்ததாக இருக்க வேண்டியதில்லை என்று அதன் வடிவமைப்பாளர்கள் செய்த வலியுறுத்தல்களே ஆகும். இவ்வாறு, எந்த நடவடிக்கையும் நிரலாக்கம் செய்யும் கருவிகளை மொழியியலாளர்களுக்கு வினியோகிப்பதற்குப் பதிலாக ஏரியன்-78 ஒவ்வொரு நிலைக்கும் விரும்பப்பட்ட மொழியியல் நடவடிக்கையை விளைவிக்கத் தேவையான குறைந்த அளவு திறனை மட்டும்தான் வினியோகிக்கும். இது மொழியியலாளர்கள் அதிக ஆர்வம் காரணமாக தேவையற்ற சிக்கல்களை உருவாக்கும் சாத்தியதத்தைக் குறைக்கும்; மேலும் அதிக பொதுவான திட்டத்தில் வேலைசெய்வதை விட கூடுதல் வேகத்தில் வேலைசெய்யும் மென்பொருள்களை உருவாக்க இயலச்செய்தது.

ரோப்ரா (ROBRA) துணை ஒழுங்குமுறையில் உள்ள "இலக்கணம்" உண்மையில் துணையிலக்கணங்களின் ஒரு வலைப்பின்னல் ஆகும்; அதாவது இலக்கணம் துணை இலக்கணங்களின் பயன்பாடுகளின் மாற்று வரிசைகளை மற்றும் துணை இலக்கணங்கள் பயன்படுத்தப்படவேண்டிய விருப்பத் தேர்வுகளைக் குறிப்பிட்டும் ஒரு வரைபடமாகும். எனவே மேல்நிலை இலக்கணம் பகுப்பாய்வு, மாற்றல், போன்ற மொழியியல் நடைமுறைகளை

உண்மையில் விளைவிக்கும் துணை இலக்கணங்களுக்கு மேலுள்ள ஒரு "கட்டுப்பாட்டு வரைபடமாகும்". ஏரியன்-78 எந்த மொழியியல் கோட்பாடுகளின் அல்லது விருப்பப்பட்டால் பல்கிய கோட்பாடுகளின் நடைமுறைப்படுத்தலையும் அனுமதிக்கும் படிக்கு போதுமான அளவு பொதுமையானதாகும். இவ்வாறு கொள்கை அடிப்படையில் அது முழுவதும் திறந்த நிலையிலுள்ளது; மற்றும் தன்னிச்சையான பொருண்மையியல் செயற்பாங்குகளுக்கும் எந்த விளக்கத்தின் "உலக மாதிரிகளுக்கு" குறிப்பையும் இடமளிக்க இயலும்.

நடைமுறையில் இந்த விளக்கம் கூடுதல் கலவைத்தன்மையானதாகும். புதிய மொழியியல் கோட்பாடுகளின் குறிப்பாக "உலக மாதிரிகளின்" அனுகூலத்தைப் பெறத் தேவையான கணினிசார் நெகிழ்வை அதிகரிக்க வேண்டி அடித்தளத்திலுள்ள மென்பொருள் பல்வேறு வழிகளில் மாற்றப்படவேண்டிவரும். துரதிருஷ்டவசமாக அடித்தளத்திலுள்ள மென்பொருள் மாற்றங்களைச் செய்வது மிகக் கடினமானதாகும் படிக்கு திடமானதாகும். இதன் விளைவாக GETA குழு எந்தத் தீவிரமான புதிய கணினிசார் உத்தியையும் பரிசோதனை செய்ய இயலாது போனது.

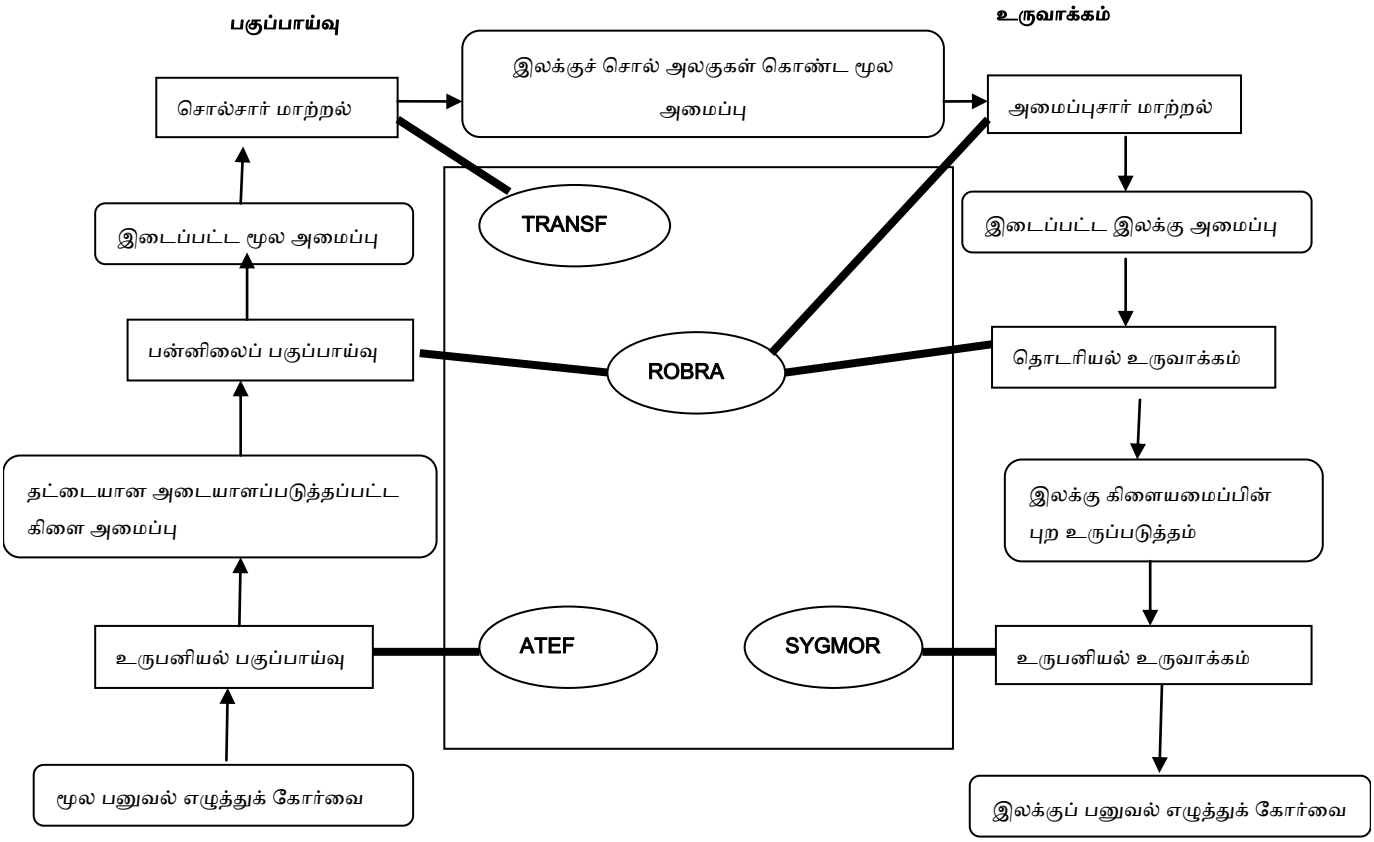
இந்த இயந்திர ஒழுங்குமுறை பல எண்ணிக்கையிலான முன்னேற்றங்களையும் மாற்றங்களையும் மேற்கொண்டுள்ளது: ஏரியன் 78, ஏரியன்- 85 மற்றும் சமீபத்திய பதிப்பு ஏரியன்-G5. இதன் நீண்ட கால இலக்கு பன்மொழிய மொழிபெயர்ப்புக்கு அடித்தளமாக இயந்திர மொழிபெயர்ப்பு 'இயந்திரம்' என்பதை உருவாக்குவதாகும். இங்கு ஏரியன்-78 விளக்கப்பட்டுள்ளது (Hutchins & Somers, 1992:221-238).

3.15.2. **பொதுவான விளக்கம்**

ஏரியன் பகுப்பாய்வு மற்றும் உருவாக்கம் இரண்டும் உருபனியல் மற்றும் தொடரியல் தொகுதிகளாகப் பிரிக்கப்பட்ட ஒரு மாற்றல் ஒழுங்குமுறையாகும் (transfer system). மாற்றலும் இரு கட்டங்களைக் கொண்டது: சொல்சார் மற்றும் அமைப்புசார் மாற்றல். ஏரியனில் பயன்படுத்தப்பட்டுள்ள உருப்படுத்த நிலைகள் மொழியியல் பார்வை அடிப்படையில் ஆர்வம் ஊட்டுவதாகும்: அவைகள் சார்பு உறவுகளையும் (dependencey relations) உறுப்பு அமைப்புகளையும் (constituent structures) இணைக்கும் பன்னிலை அமைப்புகள் (multiple-level structures) ஆகும்; மற்றும் அக மற்றும் புற/மேலோட்டமான மொழியியல் தகவல்களைக் கொண்டுள்ளன.

கீழே தந்துள்ள படம் (Hutechins and Somers, 1992:223) இந்த ஒழுங்குமுறையின் பொதுவான கட்டமைப்பைக் காட்டுகின்றது. இந்தப் படம் மூல மொழிப் பகுப்பாய்வை இடதுபக்கம் மேலும் மாற்றலை உச்சியிலும் இலக்குமொழி உருவாக்கத்தை வலதுபக்கம் கீழும்

காட்டுகின்றது. முனைமழுங்கிய செவ்வகப் பெட்டிகள் இடைப்பட்ட தரவு அமைப்புகளை உருப்படுத்தம் செய்கின்றன; செவ்வகப் பெட்டிகள் மொழியியல் செயற்பாங்குகளைப் பண்பாக்கம் செய்கின்றன. முட்டைவடிவப் பெட்டிகளில் காட்டப்பட்டுள்ள விதி எழுதும் வடிவவாதங்களை நடைமுறைப்படுத்தும் தனிப்பட்ட மென்பொருள் கருவிகள் கொண்ட உள்ளேயுள்ள பெரிய பெட்டி மென்பொருள் ஒழுகுமுறையை உருப்படுத்தம் செய்கின்றது.



மொழிபெயர்ப்பின் ஒழுக்கு ஒரு தரமான/நிலையான அடுக்குநிலையைப் (stratification) பின்பற்றுகின்றது: மூலப் பனுவல் எழுத்துக்களின் கோர்வை தட்டையான அடையாளப்படுத்தப்பட்ட கிளையமைப்பை விளைவிக்கும் படி உருபனியல் பகுப்பாய்விற்கு உள்ளாகின்றது. பன்னிலை பகுப்பாய்வு இடைப்பட்ட மூல அமைப்பைத் தருகின்றது; இது பின்னர் இருநிலை மாற்றலுக்கு உள்ளகின்றது: சொல்சார் மாற்றலில் மூலமொழிச் சொற்கள்

அல்லது சொல் அலகுகள் (lexical units) இலக்கு மொழிச் சொல் அலகுகளால் இடம்பெயர்க்கப்படுகின்றது; இதனால் மூலமொழி அமைப்பு இலக்குமொழி சொல் அலகுகளாக மாறுகின்றது. அமைப்பு மாற்றல் இடைப்பட்ட மூல அமைப்பை விளைவிக்கின்றது. இது தொடரியல் உருவாக்கத்திற்கும் உருபனியல் உருவாக்கத்திற்கும் உள்ளீடாக வரும்.

ஏரியன் நான்கு விதி எழுதும் வடிவவாதங்களைக் கொண்டிருக்கின்றது; திறம்பட்ட குறிப்பிட்ட பணிக்கான மிக உயர்நிலை நிரலாக்கமொழிகள்; ஒவ்வொன்றும் அதன் தொடர்புள்ள செயல்படுத்தத்துடன் தொடர்புடையன; அவற்றின் விவரங்கள் விதிகளை எழுதுகின்ற மொழியியலாளர்களிலிருந்து மறைக்கப்பட்டுள்ளன; கோட்பாடு அடிப்படையில் அவை மொழியியலாளர்களுக்கு ஆர்வமூட்டுவது அல்ல. இந்த நான்கு மென்பொருள் தொகுதிகளும் நான்குவேறுபட்ட தரவு அமைப்பு மற்றும் தொடர்புடைய கையாளுகைக்கு ஏற்ப நான்கு வேறுபட்ட மொழியியல் தரவாய்வை (தரவின் மீதான செயன்மைகளை) ஒத்திருக்கும். ஒவ்வொரு வடிவவாதமும் ஒதுக்கப்பட்ட பணிக்குத் தேவையான சக்தி மற்றும் நெகிழ்வுடன் சில கணினிமுறையின் வகையை எளிதாக்கும் படி சிறப்பாக வடிவமைக்கப்பட்டுள்ளதால் இது ஏரியனின் முக்கியமான பண்புக்கூறாகும்.

ATEF உருபனியல் ஆய்வுக்கு வேண்டி வடிவமைக்கப்பட்டுள்ளது; மற்றும் எழுத்துக்களின் கோர்வைகளை தட்டையான அடையாளப்படுத்தப்பட்ட கிளையமைப்பின் மேல் ஒழுங்குபடுத்தப்பட்ட பண்புக்கூறுகளின் கட்டுக்களாகச் சித்தரிக்கும். ஆறு நிலைகளில் மூன்றிற்குப் பயன்படுத்தப்படும், குறிப்பாக இரண்டு மிக முக்கியமான நிலைகளுக்குப் பயன்படுத்தபடும் ROBRA ஒரு திறன்வாய்ந்த மென்பொருள் கருவியாகும்; இது கிளையமைப்பு குறுக்குக் கடத்துகைகளின் (transductions) விளக்கத்தை அனுமதிக்கும்; அதாவது இது தன்னிச்சையான/இடுகுறித்தன்மையான கலவைத்தன்மையான அமைப்புகளை வெளியீடாக ஏற்கவும் இந்த அமைப்புகளைக் கையாளவும் புதிய கிளை அமைப்புகளை வெளியீடு செய்யவும் அனுமதிக்கும். சொல்சார் மாற்றலுக்குப் பயன்படுத்தப்படும் TRANSF சிறிது குறைந்த திறனுள்ள வடிவமாதமாகும்; இது கிளையமைப்புகளை உள்ளீடாக ஏற்கும்; ஆனால் அவற்றை ஒழுங்குபடுத்தும் திறன் இல்லை; கிளைகள் மீதுள்ள அடையாளத்தை/புலக்குறிப்பை மாற்ற மட்டுமே திறன் உள்ளது. உருபனியல் உருவாக்கத்திற்குப் பயன்படுத்தப்படும் SYGMOR புலக்குறிப்பு செய்யப்பட்ட கிளையமைப்புகளை எழுத்துக் கோர்வைகளாக மற்றும்

செயல்பாட்டைக் கொண்டுள்ளது; இது இந்த கட்டத்தின் குறைந்த கலவைத் தன்மையை பிரதிபலிக்கும் ஒரு இறுதிநிலை தீர்மானிக்கப்பட்ட தானியியங்கியை மெய்படுத்தம் செய்யும்.

3.15.3. **பன்னிலை உருப்படுத்தம்**

கெதா அணுகுமுறையின் நேர்மறையான பண்புக்கூறுகளில் ஒன்று அதன் செயல்முறைகளிலும் உருப்படுத்தங்களிலும் மொழியியல் கோட்பாட்டை இணைத்துள்ளதாகும். ஒட்டுமொத்தமான தொகுதி ஒழுங்குமுறை வடிவமைப்புக்குள் ஒவ்வொரு தொகுதியும் மொழியியல் அடிப்படையில் ஊக்குவிக்கப்பட்டதாகும். GETA குழுவால் கைக்கொள்ளப்பட்ட மொழியியல் அணுகுமுறை, பகுப்பாய்வின் வெளியீடாகவும் மற்றும் சொல்சார் மற்றும் அமைப்பு மாற்றலுக்கும் பின்னர் உருவாக்கத்திற்கான உள்ளீடாகவும் நோக்கமாகக் கொண்ட உருப்படுத்தம் ஆகும். இது இடைமுக அமைப்பாகும்; இதில் அக்கால மொழியியல் கோட்பாட்டின் தாக்கம் குறிப்பாகத் தெளிவானதாகும்.

இடைமுக அமைப்பு அடுக்கடுக்கான மொழியியல் ஆய்விலிருந்து விளைந்த மொழியியல் உருப்படுத்தமாகும். தொகுதிமயமான அணுகுமுறையில் (modular approach) எதிர்பார்க்கவியலும் ஆய்வின் பல்வேறு நிலைகள் இருக்கும். இவை மொழியியல் உருப்படுத்தத்தில் பிரதிபலிக்கும்; இது ஒரேசமயத்தில் வேறுபட்ட நிலைகளில் தகவல்களை இணைக்கும்: உருபனியல், தொடரியல், தர்க்க-பொருண்மையியல் நிலைகள். இந்த பன்னிலை உருப்படுத்த அமைப்புத் தான் முக்கியமாகக் குறிப்பிடத்தகுந்ததாகும் மற்றும் மிகச் செல்வாக்குள்ளதாகும்.

உருபனியல் நிலையில் உருப்படுத்தம் *dogs* போன்ற பனுவல் கோர்வைகளுக்கும் அவற்றின் உருபன்சார்-சொல்சார் விளக்களுக்கும் இடையிலான வேறுபாடுகளைக் கண்டுபிடிக்கின்றது. எ.கா சொல் அலகு *dog,* பன்மைப் பெயர். தொடரியல் நிலை (புற) தொடரியல் உறவுகளான எழுவாய் மற்றும் செயப்படுபொருள், நிரப்பி மற்றும் அடை இவற்றுடன் பெயர்த்தொடர் மற்றும் வினைக் குழு போன்ற தொடரமைப்பு உறுப்புகள் அடிப்படையில் அமையும் உருப்படுத்தம் ஆகும். தருக்க-பொருண்மையியல் நிலை தருக்க இயல்பின் சார்பு உறவுகளைக் (தலைகள் மற்றும் அடைகள்) காட்டும் அகத் தொடரியல் உருப்படுத்தம் ஆகும்,

எ.கா. பயனிலைகள் மற்றும் பங்கேற்பாளர்கள் (arguments) அல்லது (இலக்கு, காரணம், இடம் போன்ற) பொருண்மைப் பங்குகளுடன் கூடிய சூழ்நிலைத் தனிமங்கள்.

3.15.4. மொழியியல் செயற்பாங்குகள்

3.15.4.1. உருபனியல் பகுப்பாய்வு

ஒரு இடைமுக உருப்படுத்ததை எட்டுவதற்கான மொழியியல் செயற்பாங்குகள் ஏறக்குறைய பின்வருமாறு அமையும். முதலாவது ATEF வடிவவாத்தில் எழுதப்பட்ட விதிகளைப் பயன்படுத்தி பனுவல் உருபனியல் பகுப்பாய்வு கட்டதிற்கு உள்ளாக்கப்படும். இக்கட்டம் அகராதி பார்த்தல் மற்றும் கோர்வை பிரித்தல் இவற்றின் இணைப்பு அடிப்படையில் ஒவ்வொரு இறுதிக் கணுவிற்கும் புலக்குறிப்பைத் தரும். பெரும்பான்மையும் உருபனியல் பகுப்பாய்வின் விளைவு பொருண்மை மயக்கம் உள்ளதாய் இருக்கும்.

உருபனியல் பகுப்பாய்வின் வெளியீடு தட்டையான புலக்குறிப்பு செய்யப்பட்ட கிளையமைப்புகள் ஆகும்; இதில் ஒவ்வொரு குறைந்த கிளையமைப்பும் வாக்கியத்தில் எத்தனைச் சொற்கள் உள்ளனவோ அதேபோன்ற எண்ணிக்கையில் உள்ள குழந்தைக் கணுக்களை ஆதிக்கம் செய்யும் ஒரு போலி பெற்றோர் கணுவைக் கொண்டிருக்கும்; இங்கு இடை உறுப்புகள் இருப்பதில்லை. ஒவ்வொரு தட்டையான கிளையமைப்பும் பொருள்மயக்கம் உள்ள ஏதாவது ஒன்றின் அர்த்ததைக் கொண்டிருக்கும்; இதனால் தனிச் சொற்களின் சாத்தியமான உருபனியல் பகுப்பாய்வின் இணைப்புகளை ஒக்கும் எண்ணிக்கையிலான வெளியீடுகள் இருக்கும்.

3.15.4.2. பன்னிலைப் பகுபாய்வு

அடுத்த கட்டம் பன்னிலை பகுப்பாய்வு ஆகும். இது இந்த ஒழுங்குமுறையின் கலவைத்தன்மையான பகுதியாகும்; இது திறன் வாய்ந்த ROBRA வடிவவாதத்தைப் பயன்படுத்துகின்றது. இது தொடக்க தொடரியல் கிளையமைப்பை உருவாக்கப் பகுப்பாய்வையும் புற தொடரியல் கிளையமைப்புகள் அக உருப்படுத்தங்களாகச் சித்தரிக்கப்படும் ஒருவகையிலான 'மாற்று' நிலையையும் இணைப்பதை உள்ளடக்கும். இதில் தொடரா சொல் அலகுகள் 'சேர்க்கப்படும்' மற்றும் பொருட்கள் (objects) தர்க்க-பொருண்மையியல் உறவுகளால் இடம் பெயர்க்கப்படும்.

"பகுப்பாய்வு' கட்டம் தொடர்ச்சியான சொல் அலகுகளை ஏற்று இடை உறுப்புகளை உருவாக்கும் விதிகளின் பிரயோகத்தை உள்ளடக்கும்.

3.15.4.3. **மாற்றல் மற்றும் உருவாக்கம்**

பன்னிலை ஆய்வின் வெளியீடு/விடுபொருள் சார்பு மற்றும் தொடரமைப்பு உறவுகளைக் காட்டும் மற்றும் தர்க்க-பொருண்மையியல் மற்றும் புறத் தொடரியல் செயற்பாடுகளைக் குறிப்பிடும் ஒரு கிளை அமைப்பாகும். மொழிபெயர்ச்சி செயற்பாங்கின் அடுத்த நிலை சொலார் மாற்றல் ஆகும்; இதில் மூல மொழி சொல் அலகுகள் தொடர்புடைய இலக்கு மொழிச் சொற்களால் இடம்பெயர்க்கப்படும்.

இந்த அமைப்பு அமைப்புச் சார் மாற்றலுக்குக் கடத்தப்படும்; அது தேவையான அமைப்புச்சார் மாற்றங்களைச் செய்யும்; இது பங்கேற்பாளர்களுக்கு மறு எண்ணிடுதல், வேற்றுமை அடையாளத்தை தருவது அல்லது நீக்குவது, பெரும்பான்மையும் கிளையமைப்பை மறுசீரமைப்பு செய்வது போன்றவற்றை உட்படுத்தும். உள்ளீடு இலக்குமொழி அமைப்புகளாகும் (இலக்குமொழி சொல் அலகுகள் ஏற்கனவே செருகப்பட்டுள்ளது); வெளீடு இலக்குமொழியின் (அக) அமைப்புகளுடன் ஒத்திருக்கவேண்டும். கிளையமைப்பு குறுக்குகடத்துகை வடிவவாதம் (transduction formalism) ROBRA பங்கேற்பாளர் குறியீட்டல் மாற்றம் மற்றும் கூடுதல் கலவைத்தன்மையான கிளையமைப்பு மாற்றங்கள் என்ற கையாளுகைக்கு ஆளாகும்.

உருவாக்கத்தில் அடுத்த நிலையான 'தொடரியல் உருவாக்கம்' புற அமைப்பு புலக்குறிப்புகள் ஒதுக்குவதை உள்ளடக்கும்; அதாவது புற எழுவாய்கள் மற்றும் செயப்படுபொருள்களின் தெரிவு, பொருத்தமான துணைவினைகளின் தேர்வு போன்றன, சொல் வரிசையின் மறு ஒழுங்குபடுத்தல், உருபனியல் மாறிகளுக்கு மதிப்புகளைத் தீர்மானித்தல் (எ.கா. எண் மற்றும் பால் உடன்பாடு). இது மீண்டும் கிளையமைப்புக்களின் மறு அமைப்பை உள்ளடக்கும்; மற்றும் ROBRA வின் வடிவவாதத்தின் பயன்பாடு.

உருபனியல் உருவாக்கம் மொழிபெயர்ப்புச் செயற்பாங்கின் இறுதி நிலையாகும். SYGMOR என்ற தொடர்புடைய விதி எழுதும் வடிவவாதம் தொடரியல் உருவாக்கக் கட்டத்தில் புலக்குறிப்புச் செய்யப்பட்ட கிளையமைப்புகளை நிறுத்தற்குறிகளை உள்ளடக்கிய எழுத்துக் கோர்வைகளாக மாற்றும் செயற்பாட்டைக் கொண்டிருக்கும். முதலாவது கிளையமைப்புகள் மதிப்புகள் மற்றும் பண்புகளின் இணைகளின் கோர்வைகளாக மாற்றப்படும்.

3.15.5. **விதி எழுதும் வடிவவாதங்கள்**

முன்னர் கூறியபடி ஏரியனின் நான்கு மென்பொருள் தொகுப்புகள் மொழியியல் தரவு ஆய்வின் நான்கு வேறுபட்ட வகைகளுடன் ஒத்திருக்கும்: கோர்வைகளிலிருந்து கிளைகள் (ATEF), கிளையமைப்புகளிலிருந்து கிளைஅமைப்புகள் (ROBRA மற்றும் TRANSF) மற்றும் கிளையமைப்புகளிலிருந்து கோர்வைகள் (SYGMOR).

ஏரியனின் அடித்தளத்திலிருக்கும் கணினிசார் அமைப்பு ஒரு 'உற்பத்தி ஒழுங்குமுறை' ('production system') ஆகும். அறிவித்தல் அம்சம் (declarative aspect) நடைமுறை அம்சத்திலிருந்து (procedural aspect) பிரிக்கப்பட்டுள்ளது; அதாவது வெளிப்படுத்தப்படவேண்டிய மொழியியல் அறிவு எவ்வாறு மொழியியல் அறிவு பயன்படுத்தப்படவேண்டும் என்ற அறிவிலிருந்து பிரிக்கப்பட்டுள்ளது. உற்பத்தி ஒழுங்குமுறை தரவு அமைப்பு (data structure), விதி தொகுப்பு (rule set), விளக்கி (interpreter) என்பனவற்றைக் கொண்டிருக்கும் எனக் கூறப்பட்டது. இந்நேர்வில் தரவு அமைப்பு கிளையமைப்புகளைக் குறிப்பிடும் புலக்குறிப்பு விற்களைக் கொண்ட வரைபடங்களாகச் செயல்படுத்தப்பட்டுள்ள மொழியியல் உருப்படுத்தம் ஆகும். விதித் தொகுப்பு இந்த வடிவவாதத்தில் எழுதப்பட்டுள்ள மொழியியல் விதிகளின் தொகுப்பாகும். விளக்கி விதி எழுதும் வடிவவாதத்தை செயல்படுத்தும் அடித்தளத்திலுள்ள கணினி நிரல் ஆகும்.

3.15.5. முடிவுரை

கெதா நவீன வகை 'இரண்டாவது தலைமுறை' மொழிபெயர்ப்பு ஒழுங்குமுறை ஆகும்; அதாவது மொழியியல் சார்புடைய அதிக தொகுதிமயமான மறைமுகமான மொழிபெயர்ப்பு ஒழுங்குமுறையாகும்; வழிமுறைசார் மற்றும் மொழியியல்சார் செயற்பாங்குகளைத் தெளிவாகப் பிரிக்கும் பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் உரிய அடுகப்பட்ட அணுகுமுறையாகும்; இதில் பொருண்மையியல் ஆய்வு இல்லை; செயற்கை அறிவு 'புரிந்துகொள்கை' இல்லை; ஊடாடும் வசதிகள் இல்லை.

3.16. யுரொட்ரா (EUROTRA)

ஹட்சின்ஸ் மற்றும் சோமர்ஸ் (Hutchins and Somers 1992) மற்றும் சோமர்ஸ் (1985) யுரொட்ரா பற்றி எழுதிய தகவல்களின் அடிப்படையில் இப்பகுதி அமையும்.

3.16.1. பின்னணி

யூரோட்ரா என்பது 1978 முதல் 1992 வரை ஐரோப்பிய ஆணையத்தால் (European Commision) நிறுவப்பட்டு நிதியளிக்கப்பட்ட ஒரு லட்சிய இயந்திர மொழிபெயர்ப்பு திட்டமாகும். யுரொட்ரா மேற்கத்திய உலகில் மிகப்பெரிய இயந்திர மொழிபெயர்ப்பு ஆய்வுத்திட்டமாகும். ஒரு உண்மையான பன்மொழிய மொழிபெயர்ப்புத் திட்டம் உருவாக்க எடுக்கப்பட்ட முதல் முக்கியமான முயற்சியாகும்; இது ஏழு ஐரோப்பிய பொருளாதாரக் கூட்டமைப்பு (European Economic Community (EEC)) மொழிகளை உட்படுத்தியதாகும். இந்தத் திட்டத்திற்கான நியாயம் தவிர்க்க இயலாத பொருளாதாரம் ஆகும். ஐரோப்பிய பொருளாதாரக் கூட்டமைப்பின் முழு நிர்வாக வரவுசெலவுத்திட்டத்தின் மூன்றில் ஒரு பகுதிக்கு மேல் மொழிபெயர்ப்புப் பிரிவுக்குக் கொடுக்கத் தேவைப்பட்டது.

யுரொட்ரா ஒரு உண்மையான பன்னாட்டு வளர்ச்சி ஆய்வுத்திட்டமாகும். திட்டம் நடத்துவதற்கு ஒரு மைய ஆய்வுக்கூடம் இல்லை; அங்க நாடுகளின் நியமிக்கப்பட்ட பல்கலைக்கழப் பிரதிநிதிகள் அவரவர் மொழிகளுக்குப் பகுப்பாய்வு, உருவக்கத் தொகுதிகளை உருவாக்குகினர். மாற்றல் தொகுதி மட்டும்தான் "மையக்" குழுவால் உருவாக்கப்படும். மாற்றல் தொகுதி எவ்வளவு சிறிதாக இருக்க இயலுமோ அந்த அளவு சிறிதாய் இருக்க வடிவமைக்கப்பட்டுள்ளது; சொல் இடம்பெயர்த்தலிருந்து சிறிது கூடுதலானதைக் கொண்டிருக்கும். மென்பொருள் உருவாக்கம் மொழியியல் விதி உருவாக்கத்திலிருந்து முற்றிலும் பிரிக்கப்பட்டுள்ளது. பல ஒருங்கிணைப்பு குழுக்கள் பல்வேறு மொழிகளில் செயல்படுகின்றன மற்றும் குழுக்களுடன் ஒத்துழைப்பை வலியுறுத்துகின்றனர். யுரோட்ராவின் மொழியியல் கோட்பாட்டு அடிப்படை புதியது அல்ல. "பொருண்மையை" உருப்படுத்தம் செய்யும் அடிப்படை அமைப்புகள் சார்புக் கிளையமைப்புகளாகும் (dependency trees); அவை ஓரளவுக்கு இலக்கணத்தை எழுதுகின்ற மொழி குழுக்களின் விருப்பப்படியும் ஓரளவுக்கு மொழிக் குழுக்களின் பரஸ்பர ஒப்பந்தத்தால் கட்டுப்படுத்தப்பட்டும் பண்புக்கூறு-மதிப்பு இணைகளால் (feature-value pairs) அடையாளப்படுத்தப்பட்டுள்ளன.

ஒரு அர்த்ததில் யுரொட்ராவின் மென்பொருள் அடிப்படை புதியதாய் இல்லை. அடிப்படை விதி விளக்குவான், இலக்கணங்கள்/செயற்பாங்குகள் மீது கட்டுப்பாட்டு மொழி கொண்ட பொதுவான விரித்தெழுது ஒழுங்குமுறையாகும். ஏரியன்-78-இல் உள்ளது போன்று மொழியியல் விதிகளைத் துணை இலக்கணங்களின் தொகுப்புகளாகக் கட்ட இயலும்; மற்றும் மொழியியலாளர்களுக்கு எந்தத் தொகுப்பு விதிகளைப் பயன்படுத்தப்படவேண்டும் எப்போது பயன்படுத்தப்படவேண்டும் என்பதைக் கட்டுப்படுத்த வழிகள் தரப்பட்டுள்ளன; தனிப்பட்ட விதிகள் அழிக்க இயலாத விரித்தொழுது விதிகளாகும்; இதன் காரணமாக எந்த விதியின்

பயன்பாடும் புதிய அமைப்பை உருவாக்க இயலும்; ஆனால் அதே சமயம் எந்தப் பழைய தகவலையும் அழிக்காது.

மற்றொரு அர்த்ததில் யுரொட்ராவின் மென்பொருள் அடிப்படை அதற்கு முந்தைய பிற ஒழுங்குமுறைகளிலிருந்து சொல்லத்தக்க அளவுக்கு வேறுபட்டதாகும். பகுப்பாய்வு, மாற்றல், மற்றும் உருவாக்க உபாயங்கள் நிரலர்கள் நடைமுறைப்படுத்தும் வழிமுறைகளில் உட்படுத்தப்படவில்லை. மாறாக அவை மொழியியலாளர்களால் முறைப்படுத்தப்பட்டு ஒரு தனித்தன்மையான கட்டுப்பாட்டு மொழியால் உருப்படுத்தம் செய்யப்பட்டுள்ளது.

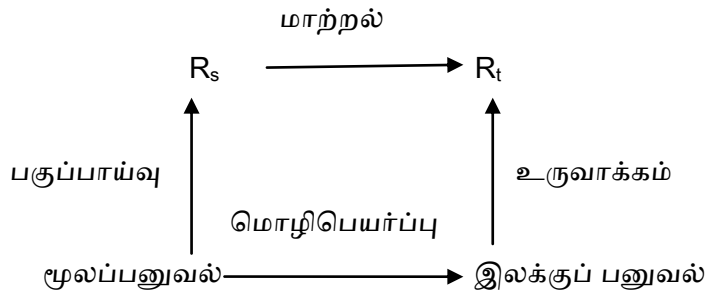
3.16.2. **அமைப்பு முறையும் ஒழுங்குமுறை வடிவமைப்பும்**

பகிர்ந்தாளுதல் மற்றும் பன்மொழியம் இவற்றின் தேவைகள் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையின் உருவாக்கத்தையும் வடிவமைப்பு அடிப்படையையும் தீர்மானித்தது. யுரெட்ரா உண்மையிலேயே பன்மொழிய இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்குவதின் முதல் முயற்சி எனலாம். 1978-இல் எடுக்கப்பட்ட முக்கியமான தீர்மானம் யுரொட்ரா மாற்றல்-அடைப்படையான ஒழுங்குமுறையாகும் என்பதாகும். இருப்பினும் 30, 42, 75 எனப் படிப்படியாக இணைமொழிகளை உட்படுத்தவேண்டும் என்ற நிலையில் இடைமொழி அணுகுமுறையைத் தேர்ந்தெடுப்பது உசித்தமாய் இருந்திருக்கும். இருப்பினும் அக்காலகட்டத்தில் இடைமொழி அணுகுமுறை மதிப்பிழந்திருந்தது காரணமாக அது சாத்திமில்லாது போனது. இடைமுக அமைப்புக்கள், முக்கிய விதிமுறைகள், மென்பொருள் இவற்றிற்கான அடிப்படை விவரக்குறிப்புகள் ஐரோப்பிய சமூகம் முழுவதிலும் உள்ள உறுப்பினர்களைக்கொண்ட அணிகளால் உருவாக்கப்பட்டன. தனிப்பட்ட ஆய்வுக் குழுக்கள் அவரவர் மொழிகளுக்கு பகுப்பாய்வு மற்றும் உருவாக்கத் தொகுதிகள் இவற்றின் ஆக்கத்தில் ஈடுபட்டன; 72 மாற்றல் தொகுதிகளுகளின் ஆக்கத்தில் ஈரிணைகளாக ஈடுபட்டன.

இடைமுக அமைப்பின் வரையறை விளக்கம் மிக முக்கியமானதாகும். கோட்பாடு அடிப்படையில் இது மாற்றல் அடிப்படையிலான ஒழுங்குமுறையாதலால் ஒரு குறிப்பிட்ட மொழியின் இடைமுக அமைப்பு ஒரு மொழி ஈரிணைக்கும் மற்றொரு மொழி ஈரிணைக்கும் இடையில் வேறுபடும்; எடுத்துக்காட்டாக பிரஞ்சுமொழியை ஜெர்மானிய மொழிக்கு மொழிபெயர்க்கும் போது ஒரு இடைமுக அமைப்பு இருக்கக்கூடும்; பிரஞ்சுமொழியை ஸ்பானிஷ் மொழிக்கு மொழிபெயர்க்கும் போது வேறு ஒரு இடைமுக அமைப்பு இருக்கக்கூடும். இருப்பினும் இது குறந்த அளவே நடைமுறை சாத்தியமானது; ஆனால் வெளிப்படையாக இது இலக்கு மொழி

மற்றும் மூல மொழி அடிப்படையில் நடுநிலையாக இருக்கும் ஒருமொழிய பகுப்பாய்வு (monolingual analysis) மற்றும் உருவாக்க உட்கூறுகளைக் (generation components) கொண்டிருக்கும் படியான நோக்கத்தை கீழறுக்கக்கூடும். இந்தக் காரணத்திற்காக எவ்வாறு இடைமுக அமைப்புகள் இருக்கவேண்டும் மற்றும் மொழியியல் தகவலின் எந்த நிலையிலான அருவமாக்கத்தை அது கொண்டிருக்கவேண்டும் என்பதுகுறித்து ஒரு பொதுவான கொள்கைகளின் தொகுதி ஒப்புக்கொள்ளப்பட்டுள்ளது.

யுரொட்ரா மொழிபெயர்ப்பு மூலமொழிப் பனுவலை/உரையை இலக்குமொழிப் பனுவலுக்கு/உரைக்கு பொருத்துவதாகக் கருத்தப்படுகின்றது; இங்கு பொருத்துதல் ஒழுங்குமுறையின் தொகுதிகளுடன் தொடர்புடைய நடவடிக்கைகளாக வரையறை விளக்கம் செய்யப்பட்டுள்ளது. இந்த மூன்று அடிப்படை நிலைகளில், "பகுப்பாய்வு" மூலப் பனுவலிருந்து மூலப் பனுவலைப் பிரதிபலிக்கும் ஒரு உருப்படுத்தத்திற்குப் (R_s) பொருத்துவதாகும்; "மாற்றல்" R_s-ஐ இலக்குமொழியைப் பிரதிபலிக்கும் உருப்படுத்தத்திற்குப் (R_t) பொருத்துவதாகும்; "உருவாக்கம்" என்பது R_t-ஐ இலக்குப் பனுவலாகப் பொருத்துவதாகும்.



பொருத்தத்தின் கருத்துச்சாயல்

72 மொழி ஈரிணைகளுக்கான ஒழுங்குமுறையில் மாற்றலை எளிமையாக வைத்திருப்பது வெளிப்படையான தேவையாகும்; ஆனால் இதன் விளைவாக, பகுப்பாய்வு மற்றும் உருவாகத் தொகுதிகள் அதற்கேற்ப பிற மாற்றல் அடிப்படையான ஒழுங்குமுறைகளைக் காட்டிலும் கலவைத்தன்மையானதாகும். இந்தக் கலவைத்தன்மை காரணமாக பகுப்பாய்வும் உருவாக்கமும்

பொருத்தங்களின் தொடர்ச்சிகளாகப் பிரிக்கப்பட்டுள்ளன; இது யுரொட்ரா கட்டமைப்புக்கு அதன் அடிப்படையான அடுக்கமைவுத் தன்மையைத் தருகின்றது. அண்மைப்படும் உருப்படுத்தங்கள் மொழியியல் அடிப்படையில் ஊக்குவிக்கப்பட்டதாகும்.

படத்தின் உச்சத்தில் உள்ள உருப்படுத்தங்கள் அதாவது பகுப்பாய்வின் இறுதி உருப்படுத்தமும் உருவாக்கத்தின் தொடக்க உருப்படுத்தமும் முன்னர் கூறிய இடைமுக அமைப்புகளாகும். இரு பக்கமும் உருப்படுத்தத்தின் நிலைகள் ஒரே எண்ணிக்கையில் இருக்கவேண்டும் என்ற நிபந்தனை இல்லை. இந்த நிலைகள் பொதுவாக மொழியியல் நிகழ்வின் உருபனியல், புறவமைப்புத் தொடரியல் (உறுப்பு), உறவுசார் மற்றும் அகவமைப்புத் தொடரியல் ஆகிய வருணனைகளின் தெரிந்த வேறுபாடுகளைப் பிரதிபலிக்கும். அவைகள் பின்வருமாறு சித்தரிக்கப் பட்டுள்ளனன.

யுரொட்ரா பனுவல் அமைப்பு: பனுவல் அல்லாத தரவு மற்றும் விளக்கப்படங்கள் உள்ளடங்கிய அமைப்புருவாக்கம் மற்றும் வெளியீட்டு குறியங்களுடன் பெறப்படும் உள்ளீடு

யுரொட்ரா இயல்பாக்கப்பட்ட பனுவல்: கிட்டத்தட்ட ASCII கோப்பை ஒக்கும் எல்லாப் பனுவல் அல்லாத தரவு மற்றும் குறியம் நீக்கப்பட்ட உள்ளீட்டுத் தரவு

யுரொட்ரா உருபனியல் அமைப்பு: அடையாளப்படுத்தப்பட்ட சொல் கிளை அமைப்புகள் மற்றும் நிறுத்தற் குறிகளின் தொடர்ச்சி வடிவிலுள்ள சொற்கள் மற்றும் உருபன்களின் உருப்படுத்தம்.

யுரொட்ரா உறுப்பு அமைப்பு: தொடரியல் வகைப்பாடுகளின் உறவுகள் அடிப்படையில் அமைந்த புறத் தொடரியல் உறுப்பு அமைப்பின் உருப்படுத்தம்.

யுரொட்ரா உறவுசார் அமைப்பு: பரந்த அளவில் சொல்செயல்பாட்டு இலக்கணத்தின் (Lexical funcational grammar) f-அமைப்புகளை ஒக்கும் தொடரியல் வகைப்பாடுகளுடன் ஒன்றாக இணைந்து (எழுவாய், செயப்படுபொருள் போன்ற) புறவமைப்பு இலக்கண உறவுகளின் உருப்படுத்தம்.

இடைமுக அமைப்பு: வேற்றுமை சட்டக அமைப்புகள் மற்றும் (விலங்கின, மானிட போன்ற) பொருண்மையியல் பண்புக்கூறுகளை உள்ளடக்கிய பொருண்மை சார்பு அடிப்படையிலான உருப்படுத்தம். இடைமுக அமைப்பு உருப்படுத்தங்கள் மாற்றல் உட்கூறுகளுக்கு உள்ளீடாகவும் அதிலிருந்து வெளியீடாகவும் வரும்.

3.16.3. கணினி அணுகுமுறை

யுரெட்ரா திட்டத்தில் இரண்டு நிலைகளின் பொருத்தத்தின் உருப்படுத்தம் மொழிபெயர்ப்பு எனப்படும். யுரெட்ராவின் விதி-எழுதும் முறைமை இ-சட்டகம் (E-framework) எனப்படும். இ-சட்டகம் என்பது ஒருங்கிணைப்பு அடிப்படை இலக்கண முறைமை (unification-based grammar formalism) என்பதில் உட்படும்.

3.16.3.1. பொருட்கள்மற்றும் அமைப்புகள் (objects and structures)

ஒவ்வொரு நிலையிலும் உருப்படுத்தம் பழம் பொருட்களையும் (primitive objects) இந்தப் பழமையான பொருட்களிலிருந்து கட்டப்பட்ட அமைப்புகளையும் கொண்டிருக்கும். இந்த பழம் பொருட்கள் பண்புக்கூறுகளின் கட்டுகளாகும் (feature bundles); பண்புக்கூறுகள் அடை-மதிப்பு இணைகளாகும்; ஒவ்வொரு நிலை உருப்படுத்தத்திற்கும் அடைகளின் சாத்தியாமான சேர்க்கைகள் பண்புக்கூறு கோட்பாட்டால் விளக்கப்பட்டிருக்கும்.

3.16.3.2. மொழிபெயப்பான்கள் மற்றும் உருவாக்கிகள் (Translators and generators)

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு ஆர்வமூட்டும் பொருட்கள் தொகுக்கப்பட்ட பொருட்கள் எனப்படும்; இதன் அர்த்தம் இவ்வெலா விதிகளாலும் இவை நல்லுருவாக்கம் செய்யப்பட்டவைகளாகும். தொகுக்கப்படாத பொருட்களும் உண்டு; அவை நல்லுருவாகம் செய்யப்பட்ட அல்லது செய்யப்படாத அமைப்புகளாகும்; அவை கருதப்பட்ட அமைப்புகளாகும் (hypothetical structures); அவற்றின் மதிப்பு நிரூபிக்கப்படவோ மறுக்கப்படவோ செய்யப்படுவதற்காகக் காத்திருக்கின்றன. அவைகள் பகுதி குறிப்பிடப்பட்ட அல்லது சீரற்ற அல்லது முழுமை பெறாத பண்புக்கூறு விவரக்குறிப்பாகும் அல்லது மறுவமைப்பட தேவையனதாகும்.

உருவாக்கி தொகுக்கும் வேலையை பல வழிகளில் செய்யும். மூன்று முக்கிய வகையான விதிகள் உருவாக்கியில் காணப்படும். ஒவ்வொன்றிற்கும் இரண்டு மாறிகள் உள்ளன.

மொழிபெயர்ப்பிகள் தொகுக்கப்பட்ட பொருட்களை ஒரு நிலையிலிருந்து மற்றொரு நிலக்கு தொகுக்கப்படாத பொருட்களின் ஒரு குழுமாக பொருத்தம் செய்யும். இரண்டு அடிப்படையிலான கொள்கைகள் மொழிபெயர்ப்பிகளின் விளக்கத்தை வழிகாட்ட உருவாக்கப்பட்டுள்ள. ஒவ்வொரு மொழிபெயர்ப்பிகளும் இருவகை டி-விதிகளின் குழுமத்தைக் கொண்டிருக்கும்.

3.16.3.3. நடைமுறைபடுத்தல் (Implementation)

யுரெட்ரா சமூகத்தின் கலைச்சொல்லில் 'நடைமுறைபடுத்தல்' என்பது அடிப்படை மென்பொருளை நிறுவுதல் என்று பொருள்படாது; மொழிபெயர்ப்பு இணைகளின் உருவக்கத்தைக் குறிப்பிடும்.

3.16.4. முடிவுரை

யுரெட்ரா இக்காலக்கட்டத்தின் மிக ஆர்வமிக்க இயந்திரமொழிபெயர்ப்பு ஆய்வுத்திட்டமாகும். இதுவரை பெரிய ஆய்வு முயற்சிகள் 'அறிவியல் முன்-வளச்சி' ஒழுங்குமுறையில் கூட விளைந்ததில்லை; ஆனால் இது முன்னோக்காகப் பார்க்கப்பட வேண்டும். இந்த ஆய்வுத்திட்டம் தொடங்கப்பட்ட போது ஒரு பன்மொழிய இயந்திர மொழிபெயர்ப்புக்குத் தேவையான மொழியியல் அறிவு இருக்கவில்லை. மொழியியல் ஆய்வு தேவையான விளக்கங்களைத் தரவில்லை; பிற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் ஒரு மேம்பட்ட பன்மொழிய வடிவமைப்பால் ஏற்றுக்கொள்ளவியலும் வழிகளில் ஒழுங்குபடுத்தப்படவில்லை. இப்போது இந்த அடிப்படையான ஆய்வின் பெரும்பான்மையான பகுதி நிறைவேற்றப்பட்டுள்ளது எனக் கூறவியலும். நிறைய செய்ய வேண்டியுள்ளது. யுரொட்ரா ஐரோப்பிய மொழியியலாளர்களுக்கு எல்லா ஆய்வுத்திட்டங்களிலிருந்தும் மேம்பட்ட ஒன்றாகும்.

3.17. மெட்டல் METAL

உலகம் முழுவதும் நடைபெறும் பெரும் இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சி மற்றும் வளர்ச்சியில் ஈடுபட்டுள்ள முதன்மையான குழுக்களில் ஒன்றான டெக்சாஸ் பல்கலைக்கழக மொழியியல் ஆராய்ச்சி மையம் மெட்டல் METAL என்ற ஆய்வுத்திட்டத்தில் ஈடுபட்டுள்ளது. இது சமீபகாலத்தில் ஒரு வணிகத்ததரமான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை வழங்கியுள்ளது. மெட்டல் ஜெர்மன்-ஆங்கில ஒழுங்குமுறை பலதடவைகள் பரிசோதனை செய்யப்பட்டு திட்ட ஆதரவாளரின் இணைய தளத்தில் மேலும் பரிசோதனைக்காகவும் இறுதி உருவாக்கத்திற்கும் வேண்டி நிறுவப்பட்டுள்ளது. LITRAS என்று மறுபெயரிடப்பட்டுள்ள இந்த ஒழுங்குமுறை 1975-இல் விற்பனைக்கு விடப்பட்டுள்ளது. மெட்டல் அகராதிகள் தேர்ந்தெடுக்கப்பட்ட தொழில்நுட்பப் புலங்களின் அதிகப்படியான சாத்தியமான உள்ளடக்கப் பரப்புக்கு வேண்டி விரிவாக்கப்பட்டுள்ளது.

மெட்டல் ஒழுங்குமுறையின் குறிப்பிட்ட வலிமைகளில் ஒன்று அது பல மொழியியல் கோட்பாடுகளுக்கும் செயல்திட்டங்களுக்கும் இடமளிக்கிறது என்பதாகும். ஜெர்மன் பகுப்பாய்வுக் கூறு சூழல்கட்டுப்பாடில்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையில் அமைந்ததாகும். இது தன்னிச்சையான மாற்றங்களுக்கான (transformations) வசதிக்கு வேண்டிய செயல்முறைகளால் பெரிதாக்கப்பட்டுள்ளது. மாறாக ஆங்கிலப் பகுப்பாய்வு கூறில் மாற்றப்பட்ட பொதுமைப்படுத்தப்பட்ட சொற்றொடர் தொடரமைப்பு இலக்கணம் (generalized phrase structure grammar (GPSG)) அணுகுமுறை பயன்படுத்தப்பட்டுள்ளது; மாற்றங்களின் (transformations) உதவி பயன்படுத்தப்படவில்லை. பகுப்பாய்வு மாற்றலிருந்து (transfer) முற்றிலும் பிரிக்கப்பட்டுள்ளது. இந்த ஒழுங்குமுறை பன்மொழிய அமைப்பாகும்; இதன்படி தரப்பட்டுள்ள உறுப்பமைப்பு ஆய்வை (constituent structure analsysi) பல இலக்குமொழிகளில் மாற்றமைப்புச் செய்யப்படுவதற்கோ உருவாக்கம் செய்யப்படுவதற்கோ பயன்படுத்த இயலும்.

மெட்டலின் மாற்றமைப்புக் கூறு இரண்டு மற்ற தொகுப்புகளை உட்படுத்தியுள்ளது; ஒன்று மாற்றல் இலக்கண விதிகளாலும் இரண்டாவது மாற்றல் அகராதி பதிவுகளாலும் பயன்படுத்தப்பட்டுள்ளது. இவை மாற்றலின் போது துணைசெய்கின்றன; இது பகுப்பாய்வு நிலையில் கிளை அமைப்பு செயல்முறையின் மேலிருந்து-கீழ் செயற்பாங்கின் போது செயல்படும். மேலிருந்து-கீழ் கடத்து (top-down pass), மாற்றல் விதிகள் எழுதிய மொழியியலாளர்களால் கட்டுப்படுத்தப்பட்டுள்ளது. இவை ஆரம்ப பகுப்பாய்வை செயற்படுத்த பயன்படுத்தப்படும் இலக்கண விதிகளுடன் 1-1 என இணைசேர்க்கப்பட்டுள்ளது. எனவே பயன்படுத்தக்கூடிய விதிகளைக் கண்டுபிடிக்க பொதுவான மாற்றல் இலக்கணம் (transfer grammar) வழியாகத் தேடத் தேவையில்லை. இருப்பினும் அதிகப் பொதுவான மாற்றல் இலக்கணத்தை (general transfer grammar) பயன்படுத்தக்கூடிய விருப்பத்தேர்வு உள்ளது; இது எச்சத்தொடர்களின் (clauses) மொழிபெயர்ப்பின் போது பயன்படுத்தப்படும். முன்னர் குறிப்பால் உணர்த்தியது போன்று அமைப்புசார் மற்றும் சொல்சார் மாற்றல் (structural and lexical transfer) ஒரே கடத்தின் போது செயல்படுத்தப்படுகின்றது; இதனால் ஒன்று மற்றொன்றின் செயல்பாட்டின் போது பாதிக்க இயலும். குறிப்பாக மாற்றல் அகராதி பதிவுகள் அவை செல்லுபடியாகும் தொடரியல்சார் மற்றும்/அல்லது பொருண்மையியல் சூழல்களைக் குறிப்பிடக்கூடும். தரப்பட்டுள்ள உள்ளீட்டிற்கு எந்த பகுப்பாய்வும் கிடைக்கப்பெறாவிட்டால், மாற்றலுக்கும் உருவாக்கதிற்கும் சுதந்திரமான

மிகப்பெரிய தொடர்கள் தேர்ந்தெடுக்கப்படும்; இதனால் ஒவ்வொரு உள்ளீடும் (ஒரு வாக்கியமோ தொடரோ) ஏதாவது மொழிபெயர்ப்பைத் தரும்.

மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதுடன், டெசாஸ் குழு (அசல் உள்ளீடு ஆவணங்கள் போன்று வெளியீடு மொழிபெயர்ப்பை வடிவமைக்க) பனுவல் செயற்பாங்கு (text processing), (அகராதி உள்ளீடுகள், இலக்கண விதிகள் இவற்றின்) டேட்டா பேஸ் மேனேஜ்மெண்ட் (data base management), (அகராதிப் பதிவுகள் மற்றும் இலக்கண விதிகள் இவற்றின் பெரும்பாலான பிழைகளை குறைக்க) விதிகளைச் சரிபார்த்தல், (சொல் பதிவுகளின் குறியமாக்கலில் மனித திறனை அதிகரிக்க) அகராதி கட்டமைத்தல் முதலியவற்றிற்கு மென்பொருள் தொகுப்புகளை உருவாக்கியுள்ளது (Slocum, 1985).

3.17.1. வரலாற்றுப் பின்னணி

மெட்டலின் இயந்திர மொழிபெயர்ப்பின் ஆய்வின் மூலங்கள் டெக்ஸாஸ் பல்கலைகழகத்தில் மொழியியல் ஆய்வு மையத்தின் (Linguistic Research Centre (LRC) நிறுவுதலுக்குக் கொண்டுசெல்லும். பிற இயந்திர மொழிபெயர்ப்புக் குழுக்களைப் போல் அல்லாமல் LRCயின் வலியுறுத்தல் நீண்டகால அடிப்படை மொழியியல் ஆய்வாக அமைந்தது. ஆங்கிலம் மற்றும் ஜெர்மனின் தொடரியல் மீதான அடிப்படை ஆய்வு இருதிசை மாற்றல் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் (bidirectional transfer system) உருவாக்கத்தின் முன்மொழிபுக்கு வழிவகுத்தது.

LRCயின் இரண்டாவது கட்ட ஆய்வு 1970 இலிருந்து 1975 வரை நீட்சியடைந்தது; இது இடைமொழி அடிப்படைஅணுகுமுறை (Interlingua-based approach) ஆய்வாக அமைந்தது; தொடர்ந்து ஜெர்மன் – ஆங்கிலம் மொழிபெயர்ப்பில் கவனக்குவிப்பு செய்யப்பட்டது.

1978-இல் புதிய கட்ட ஆய்வு முனிச்-அடிப்படையிலான சீமன்ஸ் நிறுவனத்தின் நிதி உதவியுடன் தொடங்கப்பட்டது. மெட்டல் என்பது Mechanical Translation and Analysis of Language (METAL) என்பதன் சுருக்கமாகும்; மெட்டல் மொழிபெயர்ப்பு ஒழுங்குமுறை இடைமொழி அணுகுமுறையிலிருந்து மாற்றல்-அடிப்படையிலான அணுகுமுறைக்கு மாறியது.

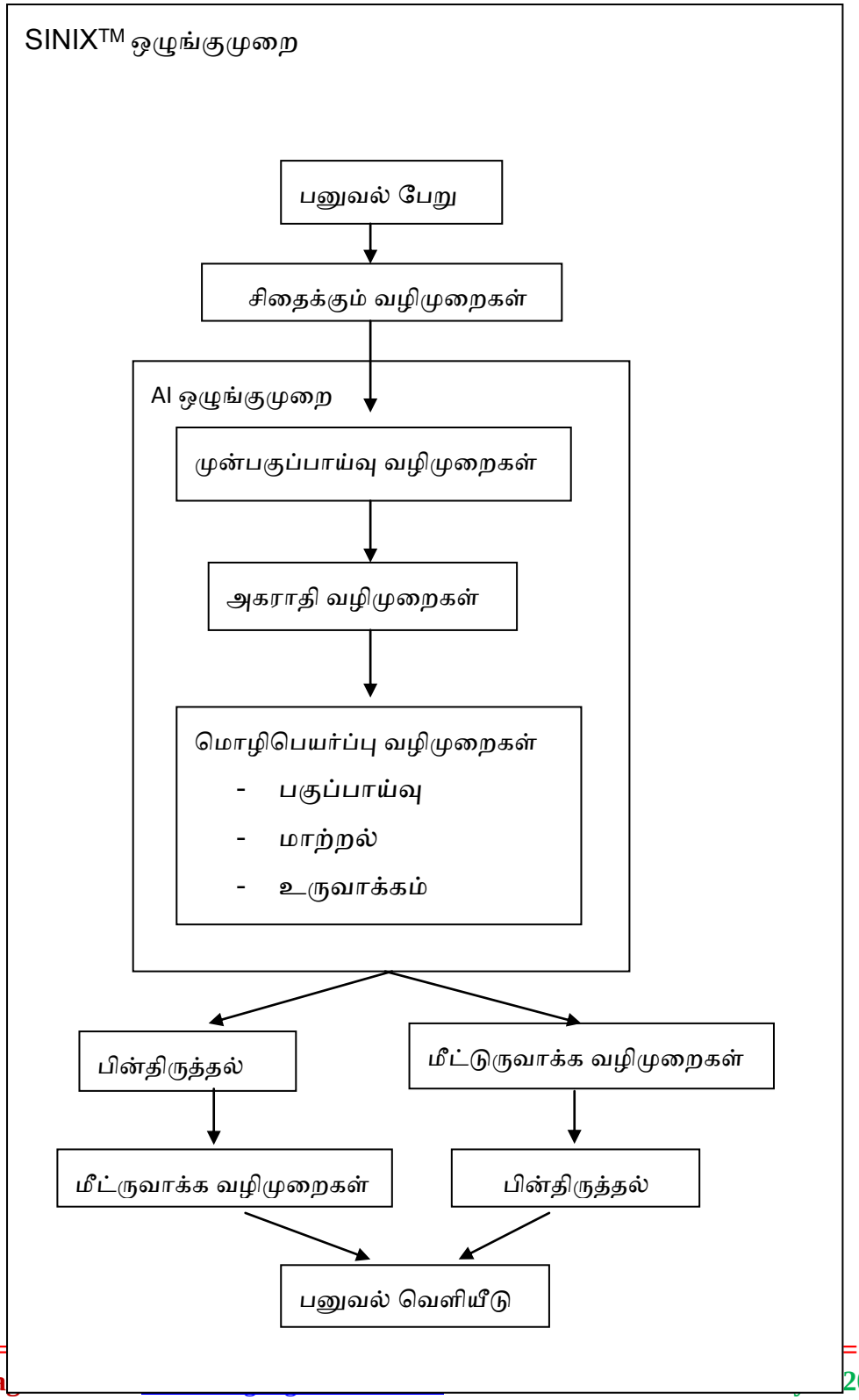
ஜெர்மன் –ஆங்கில ஒழுங்குமுறை டச்-பிரஞ்சு, ஜெர்மன் – ஸ்பானிஷ், **ஜெர்மன்-**பிரஞ்சு, ஜெர்மன் – டானிஷ் என்ற ஒழுங்குமுறைகளால் தொடரப்பட்டது.

3.17.2. அடிப்படை ஒழுங்குமுறை

மெடல் அட்டவணைகள், படங்கள், ஒழுக்கு விளகப்படங்கள், வரைபடங்கள் போன்றவைகளைக் கொண்ட பெருந்தொகுதிகளாக அமையும் தொழில் நுட்ப ஆவணங்களை கையாளும் நோக்கத்திற்காக அமைந்தது. மெட்டல் ஒழுங்கமைப்பின் முக்கிய செயற்பாங்குகள் கீழே தரப்பட்டுள்ள படத்தில் தரப்பட்டுள்ளன.

அடிப்படை ஒழுங்குமுறை

படம்: மெட்டல் முக்கியச் செயற்பாங்குகள்



1. **பனுவல்பேறு (text acquisition)**: பனுவல் பேறு தொலைத்தொடர்பு இணைப்புகள் மற்றும் பல்வேறு உள்ளீட்டு வசதிகள் இவற்றிலிருந்து பெற்றதாகும், எ.கா. காந்த நாடாக்கள் (magnetic tapes), நெகிழ் வட்டுகள் (floppy discs), ஒளிவழி எழுத்துணரிகள் (optical character recognizers)

2. **வடிவக்குலைப்பு நிரல்கள் (deformating programs)**: வடிவக்குலைப்பு நிரல்கள் விளக்கப்படங்கள், அட்டவணைகள், வரைபடங்கள் போன்றவற்றை பனுவல் தரவிலிருந்து பிரிக்கின்றன; மற்றும் பிந்தையதை மறுசெருகலுக்காக (re-insertion) அடையாளப்படுத்தும்; இந்நிலை 'மொழிபெயர்ப்பு அலகுகளின்' (translation units) (தனிச் சொல்லிலிருந்து முழு வாக்கியம் வரை) அடையாளம் காணலையும் (identification) மற்றும் அடையாளப்படுத்தலையும் (marking) உட்படுத்தும்.

3.முன் பகுப்பாய்வு (pre-analysis): அகராதித் தரவுத்தளத்தைத் தேடி மூன்று பட்டியல்களை உருவாக்குவது: (அ) அகரவரிசையில் பட்டியலிடப்பட்டுள்ள அறிப்படாத சொற்கள் (unknown words), (ஆ) அகராதியில் ஏற்கனவே உள்ள கூறுகளின் அடிப்படையில் பரிந்துரைக்கப்பட்ட சாத்தியாமான ஆங்கில மொழிபெயப்புகளைக் கொண்ட (ஜெர்மன் மொழியில் அடிக்கடி காணப்பட இயலும்) அறியப்படாத கூட்டுசொற்கள் (unkown compounds); எடுத்துக்காட்டாக *Prüfangaben* என்பது *testing specification* என்பதைப் பரிந்துரைக்கலாம்; (இ) கையிலுள்ள பனுவலுக்கு பயனாளி துல்லியத்தையும் பொருத்தத்தையும் சரிபார்க்க ஏற்கனவே அறியப்பட்ட கலைச்சொற்கள்.

4. அகராதி நிரல்கள் (dictionary programs): மூல மற்றும் இலக்கு மொழிகளின் ஒருமொழி அகராதிகள் உருபனியல், தொடரியல், பொருண்மையியல் தகவல்களைக் கொண்டிருக்கும்; இருமொழிய மாற்றல் அகராதிகள் (bilingual transfer dictionaries) மொழிபெயர்ப்பு நிகரன்களைக் குறிப்பிடும் மற்றும் சூழல் தகவல்களையும் உட்படுத்தும்.

5. **மெட்டல் மொழிபெயர்ப்பு நிரல்கள்** (METAL translation programs) (பின்னால் விளக்கப்பட்டுள்ளது)

6. **மறுவடிவாக்கம் (reformatting)**: பனுவல்சார் மற்றும் மொழியியல்சாரா தரவுகளின் இணப்பு

8. **பின்-திருத்தியமைத்தல் (post-editing):**

7. **வெளியீடு** (**output**): சொல் செயலாக்க ஒழுங்குமுறை (word processing system), அச்சு வெளியீடு (printer output), தட்டச்சிடல் (typesetting), பிற பனுவல் செயலாக்க மற்றும்/அல்லது வெளியிடும் ஒழுங்குமுறைக்கு மாற்றல்

பின்வரும் விளக்கம் முக்கியமாக 1980-களின் மையத்தில் ஜெர்மன்-ஆங்கில மொழிபெயர்ப்பு ஒழுங்குமுறையின் விவான விவரங்களின் அடிப்படையில் அமைந்த்தாகும்.

3.17.3. மொழியியல் தரவுத்தளங்கள்

மெட்டல் ஒழுங்குமுறையில் (பகுப்பாய்வு, மாற்றல், உருவாக்கம் போன்ற) மொழிபெயர்ப்பு செயன்மைளுக்கும் செயன்மைகளின் போது அழைக்கப்படுகிற தரவுக்கும் இடையில் தெளிவான பிரிவினை இருக்கின்றது. மூல மற்றும் இலக்கு மொழிகளுக்கு ஒருமொழிய சொல் தகவல்கள், ஒரு குறிப்பிட்ட மொழி இணைகளுக்கு இருமொழிய சொல் தகவல்கள், இலக்கணவிதிகளின் தொகுப்புகள் முதலியவற்றைத் தரவு கொண்டிருக்கும். பிந்தையது பல்வேறு வகையான மொழியியல் அலகுகளின் (உருபன்கள், சொற்கள், தொடர்கள், எச்சத்தொடர்கள், வாக்கியங்கள்) மீது மொழிபெயர்ப்பின் எல்லா நிலைகளிலும் பிரயோகிக்கப்படும் மற்றும் உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் பண்புகூறுகளை உள்ளடக்கும் விதிகளை உள்ளடக்கமாய்க்கொள்ளும்.

3.17.3.1. அகராதிகள்

மெட்டலின் ஒருமொழிய மூல மற்றும் இலக்கு அகராதிகள் அடிப்படை உருபனியல், தொடரியல், பொருண்மையியல் தகவல்களைக் கொண்டிருக்கும். மெட்டல் சொற்றொகுதிகளின் படிநிலையமைப்பைத் தரும்: அடிப்படைத் தொகுதிகள் (basic modules) (செயற்பாட்டுச் சொற்கள், பொது சொற்றொகுதி, பொதுவான கலைச்சொற்தொகுதி எனப்படும்) மூன்று அடிமட்ட நிலைகள் (three lowest levels) ஆகும்; இது எந்த விஷயத்தையும் பயன்படுத்தும். இந்தத் தரமான தொகுதிகளில் பயனாளிகள் எவ்வளவு சிறப்பான/தனித்தன்மையான பொருள்விளக்கச் சொற்கோவைகளையும் சேர்க்கவியலும் மற்றும் அவைகள் பயன்படுத்தப்படவேண்டிய வரிசைமுறைகளையும் குறிப்பிடவியலும். மேலும் விஷயச் சிறப்புள்ள/தனித்தன்மையுள்ள சொற்கோவைகளின் பயனாளிகள் நாட்டுச் சிறப்புள்ள/தனித்தன்மையுள்ள சொற்கோவைகளையும் விளக்க இயலும் (எடுத்துக்காட்டாக ஜெர்மன் *lastwagen* என்பது

அமெரிக்காவில் *truck* என்று மொழிபெயர்க்கப்படும்; பிரிட்டனில் *lorry* என்று மொழிபெயர்க்கப்படும்).

ஒருமொழிய அகராதிகளில் சொற்பதிவுகள் மதிப்புடன் கூடிய பண்புக்கூறுகளின் பட்டியல், எ.கா. வேர் வடிவம், இலக்கண வகைப்பாடு, உருபனியல் மாற்றுக்கள் (morphological variants), எண், இடம் போன்றன. அகராதிகள் பகுதி வடிவுகளைத் (stem forms) (பொதுவாகச் சொல்தொடக்க தனிமங்கள்) உட்படுத்துவதோடு மட்டுமல்லாமல் ஒட்டுகளையும்/விகுதிகளையும் (பொதுவாகச் சொலிறுதித் தனிமங்கள்) உட்படுத்தும்; பதிவுகள் மாற்றுப் பகுப்பாய்வுகளுக்கு இடையில் தேர்வை செயல்படுத்த 'விருப்பத்தேர்வு' மதிப்பையும் எந்த சொல் சேர்ந்துவருகைகளின் அறிகுறிகளையும் (lexical collocations) (எ.கா. தொடர்ச்சியுறா வினைவடிவுகள்: *look up*, *zurückgeben*), விஷயப் புலன்களின் அறிகுறியையும் சொற்பதிவு உட்படுத்தும். பெயர்களின் பதிவுகள் அவற்றின் திரிபு வகுப்புகளைக் காட்டும் மற்றும் உடன் வருகை கட்டுப்பாட்டிற்குப் பயன்படுத்தப்படும் பொருண்மைப் பண்கூறைச் ('இருக்கும்பொருள்/இருப்பான்', 'உயிருள்ளவை'. 'வியாபாரப்பொருள்/பண்டம்' போன்றன) சேர்க்கும்; வினைகளின் பதிவுகள் பங்கேற்பாளர்களின் (arguments) பொருண்மைப் பண்புக்கூறுகளுடன் வேற்றுமை-சட்டக விவரக்குறிப்புகள் ('அக வேற்றுமை' உறவுகள் அடிப்படையில்: செயலி, இலக்கு, பயனாளி போன்றன), பங்கேற்பாளர்களின் உறுப்பு வகை (constituent type) (எ.கா. பெயர்த்தொடர், முனுருபுத் தொடர், எச்சத்தொடர் போன்றன) மற்றும் பங்கேற்பாளர்களின் புறத் தொடரியல் செயல்பாடு (எழுவாய், செயப்படுபொருள், என்று-நிரப்பி போன்றன); அவற்றின் இணைதிறனின் (valency) விவரக்குறிப்பு ('செயப்படுபொருள் ஏற்பு வகை'): ஒரு பங்கேற்பாளர் கொண்ட செயப்படுபொருள் ஏற்காமை (செயலி), இரண்டு பங்கேற்பாளர்கள் கொண்ட செயப்படுபொருள் ஏற்காமை (செயலி, இடம்), இரண்டு பங்கெடுப்பாளர்கள் கொண்ட செய்ப்படுபொருள் ஏற்பு (செயலி, இலக்கு), போன்றன.

உள்ளீட்டுச் சொற்பதிவுகள் இலக்கணம் மற்றும் மொழிபெயர்ப்பு தகவல்களுக்காக ஊடாட்டம் அடிப்படையில் பயனாளர்களுக்கு உதவும் 'இடைக்குறியமாக்கியால்' வசதிசெய்யப்பட்டுளது. இடைக்குறியமாக்கி குறைந்த அளவு தகவலை (வேர் வடிவம் மற்றும் இலக்கண வகைப்பாடு) ஏற்கும் மற்றும் தானியக்கமாக உருபனியல் மாற்றுருக்களை உருவாக்கும் மற்றும் தொடரியல் பண்புக்கூறுகளையும் மதிப்புகளையும் குறியமாக்கும் 'சொல்சார்

இயல்புநிலை' நிரலை ('lexical default' program) உட்படுத்தும். இருமொழிய அகராதிகளில் உள்ள தகவல்கள் குறிப்பிட்ட சட்டகங்களின் விவரக்குறிப்பு, சில பொருண்மை வகையின் பங்கேற்பாளர்களின் இருப்பு, செயலி வரும் தொடர்களை (active phrases) செயலியற்ற கட்டுமானங்களாக (impersonal constructions) மாற்றுதல், தனிமங்களை சேர்த்தல் மற்றும் நீக்குதல் போன்றவற்றை உட்படுத்த இயலும். இந்தக் கருவியை ஒழுங்குமுறையை உருவாக்குபவர்களால் பயன்படுத்த இயலும்; விரும்பினால் இறுதி பயனாளிகள் தங்களது அகராதிகளை உருவாக்கவும் பயன்படுத்த இயலும்.

ஒருமொழிய அகராதிகள், மூல மொழிப் பகுப்பாய்வுக்கோ இலக்கு மொழி உருவாக்கத்திற்கோ மற்றும் எந்த மொழி சம்பந்தப்பட்டாலும் பயன்படுத்தும்படி நடுநிலையாகவும் சுதந்திரமாகவும் வடிவமைக்கப்பட்டுள்ளது. மாறாக இருமொழிய மாற்றல் அகராதிகள் ஒரு மொழிபெயர்ப்புத் திசையில் ஒரு குறிப்பிட்ட மொழி இணைக்காக வடிவமைக்கப்பட்டுள்ளது.

3.17.3.2. இலக்கண விதிகள்

மெட்டலின் இலக்கணங்கள் பரிசோதனைகள் மற்றும் கட்டுப்பாடுகளாலும் மற்றும் வெளியீட்டு அமைப்புகளின் விவரக்குறிப்புகளாலும் பெரிதாக்கப்பட்ட சூழல் கட்டுப்பாடில்லாத தொடரமைப்பு விதிகளின் வரிசையற்ற தொகுதிகளைக் கொண்டது. எல்லா விதிகளும் லிஸ்ப் (Lisp) செயல்பாடுகளாக முறைபடுத்தப்பட்டது. அவை திரிபுசார் உருபனியலையும் தொடரியல்சார் அமைப்புகளையும் உட்படுத்தும்; அவை மாற்றலின் போது செய்யப்படவேண்டிய செயல்முறைகளுடன் பகுப்பாய்வின் போது செய்யப்படவேண்டிய செயல்முறைகளைச் இணைக்கும்

இடைக்குறியமாக்கியின் (intercoder) பயன்பாட்டால் அகராதியியலாளர்களின் சொல்சார் தரவுத்தளத்தை உருவாக்கும் பணி வசதிசெய்யப்பட்டுள்ளது போல் இலக்கணத்தை எழுதுபவர்களுக்கு உதவும் படி மெட்டல்ஷாப் தொடரியல் உருவாக்கக் கருவி (Metalshop syntax development toold) என்ற ஒரு அதிநவீன மென்பொருள் ஆதரவு உள்ளது. இது மொழியியலார் மொழிபெயர்ப்பு செயற்பாங்கின் பல்வேறு நிலைகளில் அமைப்புகளை உருவாக்குவதற்கும் தரப்பட்ட அமைப்பின் ஒரு பகுதிக்கு பயன்படுத்தப்படவேண்டிய பொருத்தமான விதிகளை உடனடியாகக் கண்டறியவும் உதவுகின்றது; எடுத்துக்காட்டாக ஒரு கிளையமைப்பில் பொருத்தமான கணுவை சுட்டியால்/மவுசால் சொடுக்கிக் கண்டறியலாம். இந்த வசதி இறுதிப்

பயனாளிகளுக்கு கிடைக்காவிட்டாலும் இது கூடுதல் கட்டுப்படுத்தப்பட்ட இலக்கண உருவாக்கச் சூழலுக்கு ஆதரவு அளிக்கின்றது.

3.17.3.3. மொழிபெயர்ப்பு நிரல்கள்

மாற்றல் அடிப்படையான ஒழுங்குமுறையாக, மெட்டல் மூன்று அடிப்படையான கட்டங்களைக் கொண்டுள்ளது: பகுப்பாய்வு (analysis), மாற்றல் (transfer), உருவாக்கம் (generation). சில விவரங்களில் 'ஒருங்கிணைப்பு' (integration) என்ற நான்காவது கட்டம் பகுப்பாய்வுக்கும் உருவாகத்திற்கும் இடையில் தோன்றும்; இது வாக்கியத்திற்கு உள்ளேயுள்ள மற்றும் வாக்கியதிற்கு வெளியேயுள்ள முற்சுட்டின் சிக்கல்களை தீர்க்கும் பணியைக் கொண்டிருக்கின்றது. இந்த உட்கூறுகளுக்கு இடையிலுள்ள பிரிவுகள் முன்னால் விளக்கப்பட்ட 'சுத்தமான' வடிவத்துடன் இணங்காது. எனவே மெட்டைல் 'மாற்றப்பட்ட மாற்றல்' ஒழுங்குமுறை (modified transfer system) என்று பண்பாக்கம் செய்யப்பட்டுள்ளது.

பகுப்பாய்வின் முதல் நிலை சொல் மற்றும் உருபனியல் பகுப்பாய்வு ஆகும். உருபனியல் பகுப்பாய்வு தொடரமைப்பு இலக்கணத்தின் பிரயோகத்தால் பின் தொடரப்படும். இந்நிலையில் பொருண்மைப் பண்புக்கூறுகள் பயன்படுத்தபடுவதில்லை. தொடரியல் பகுப்பாய்வு இலக்கண விதிகளில் குறிப்பிடப்பட்டுள்ள பல்வேறு செயல்முறைகளின் உயிர்ப்பை உள்ளடக்கும் எ.கா. வேற்றுமைச் சட்டகத்தை அடையாளம் காணல், மாற்றங்களை (transformations) நிறைவேற்றுதல்.

மெட்டல் பகுப்பாய்வி பழக்கமான/பரிசயமான இணை கீழிருந்து-மேல் பாணியில் (parallel bottom-up fashion) இயங்குகிறது. இலக்கண விதிகள் 'நிலைகளாக' குழுமப்பட்டுள்ளன; பகுப்பாய்வி உயர்நிலையிலுள்ல விதிகளுக்குச் செல்லுமுன் கீழ்நிலையில் எல்லா விதிகளையும் பிரயோகிக்க முயலும். ஒன்றோ அல்லது அதற்குக் கூடுதல் விளக்கங்கள் கண்டுபிடிக்கப்படும் போது அது நிற்கும்; புற அமைப்புகளை கீழ்நிலை விதிகளைப் பயன்படுத்தி விளக்கம் காணமுடியுமென்றால் கலவைத்தன்மையான/சிக்கலான மற்றும் குறைந்த சாத்தியமான விதிகள் விட்டுவிடப்படும்; கீழ்நிலை விதிகள் வெற்றிபெறாவிட்டால் பின்னர் படிபடியாக உயர்நிலை விதிகள் முயற்சிக்கப்படும். இவ்வாறு இந்த ஒழுங்குமுறை எல்லா சாத்தியமான விளக்கங்களையும் உருவாக்குவதில்லை. உள்ளீடு முழுமையற்றதால் இருந்தால் அரைகுறையான பகுப்பாய்வும் பிரயோகிக்கப்படும்.

மாற்றல் செயல்முறைகள் (transfer procedures) மாற்றல் (transfer) என்ற அடுத்த செயல்பாட்டால் உயிர்ப்பிக்கப்படும். மையக் கட்டுப்பாடு இல்லை; செயற்பாங்கு இலக்கண விதிகளிளின் மாற்றல் பகுதிகளால் தீர்மானிக்கப்படுகின்றது. மாற்றல் (இருமொழிய அகராதியிருந்தான) சொல்சார் மாற்றல் விதிகள் மற்றும் (இலக்கணவிதிகளின் மாற்றல் பகுதிகளிலிருந்தான) அமைப்புசார் மாற்றல் விதிகளின் கலவைத்தன்மையான ஊடாட்டத்தை உள்ளடக்கும். இரண்டு வகை விதிகளும் ஒன்றையொன்று பாதிக்க இயலும். மாற்றலுக்கான மெட்டலின் உள்ளீடு ஒப்பீட்டளவில் வேற்றுமை பங்குமுறை ஒப்படைப்புகள் (case role assignments) மற்றும் சில பொருண்மை பண்புக்கூறுகள் கொண்ட ஆழமில்லா தொடரமைப்பு (shallow phrase structure) உருப்படுத்தங்களாகும். வெளியீடுகள் சொல் வரிசையின் முழு விவரக்குறிப்புகள் மற்றும் உருபனியல் உறுப்புகள் கொண்ட புற உருப்படுத்தங்களாகும். இதன் விளைவாக (உருவாக்கம் என்ற செயற்பாட்டால் உயிர்ப்பிக்கப்படும்) இறுதி உருவாக்கு நிலை (generation stage) உருபனியல் அடிப்படையில் சரியான இலக்கு மொழி கோர்வைகளை உருவாக்குவதுடன் முற்றிலும் தொடர்புடையதாகும்.

3.18.. **ரோசெட்டா** (Rosetta)

எண்ட்ஹோவெனில் உள்ள பிலிப்ஸ் ஆய்வு ஆய்வுக்கூடங்களில் மேற்கொள்ளப்பட்ட இயந்திர மொழிபெயர்ப்புத் திட்டம் அன்றைய காலகட்டத்தில் மிகப் புதுமையான பரிசோதனைத் திட்டங்களில் ஒன்றாகும் (Hutchins and Somer, 1992: 279). இதன் முக்கியமான பண்புக்கூறு மாண்டேகு இலக்கணத்தின் கொள்கைகள் அடிப்படையிலான இடைமொழி உருப்படுத்தங்களை திட்டமிடும் முயற்சி என்பதாகும். மாண்டேகு கோட்பாடு நேரடியாக தொடரியலை பொருண்மையியலோடு தொடர்புபடுத்துக்கின்றது.

3.18.1. **வரலாற்றுப் பின்னணி**

இத்திட்டம் ஃப்லிப்சில் (Philips) செய்யப்பட்ட பிலிகுவா (PHLIQA) கேள்விப்பதில் ஒழுங்குமுறை என்ற முந்தைய ஆய்வில் அதன் வேரைக் கொண்டுள்ளது. இதன் செயல்பாடு ஆங்கிலத்தில் வெளிப்படுத்தப்பட்ட கேள்விகளை தருக்க உருப்படுத்தங்களாக மாற்றுவதாகும். இது சூழல்கட்டுப்பாடில்லா இலக்கணம் அடிப்படையிலான பகுப்பானால் (parser) மேற்கொள்ளப்பட்டது; இதில் ஒவ்வொரு இலக்கண விதியும் ஒரு மொழிபெயர்ப்பு விதியுடன் இணைக்கப்பட்டு தருக்கமொழியாக இருக்கும். இருப்பினும் மொழிபெயர்ப்பு நேரடியானதல்ல: சூழல்கட்டுப்பாடிலா உருப்படுத்தங்கள் உண்மையான தருக்க உருப்படுத்தத்தைப் பெறுவதற்கும்

முன்னர் ஒரு கலப்படமான 'தருக்கம்' மற்றும் 'ஆழம்'-ஆன தொடரியல் அமைப்பால் மாற்றம் செய்யப்படும். இந்த கலப்பட அணுகுமுறையின் திருப்தி இல்லாத இயல்பு புதிய இலக்கணத்தின் வடிவமைக்குக் கொண்டு சென்றது; இது முழுவதும் கூட்டமைவானது; இதில் விதிகள் சூழல்கட்டுப்பாடிலா இலக்கணங்களை விட அதிக சக்திவாய்ந்தது. தத்துவாதியான ரிச்சர்ட் மாண்டேக் (Richard Montague) என்பவரால் விளக்கப்பட்ட இலக்கணங்கள் இந்த அணுகுமுறைக்கு கவர்ச்சியான மதிரியைத் தந்தது என்ற முடிவுக்கு வரப்பட்டது.

ஜான் லாண்ட்ஸ்பெர்கன் (Jan Landsbergen) ரொசெட்டா திட்டத்தில் சாத்தியங்களை ஆயத் தீர்மானித்தார்; இது 1980 இல் தொடங்கப்பட்டது. ஆரம்பத்தில் இரண்டு பரிசோதனை ஒழுங்குமுறைகள் உருவாக்கப்பட்டன: ரொசெட்டா 1 மற்றும் ரொசெட்டா2. 1985-இல் பெரிய திட்டம் இரண்டு கட்டங்களில் தொடங்கப்பட்து. முதலாவது கட்டம் முக்கியமான மொழியியல் மற்றும் கணினியல் சட்டக்கத்திலும் டச்சிலிருந்து ஆங்கிலம் மற்றும் ஸ்ஃபானிஷுக்கு மற்றும் ஆங்கிலம் அல்லது ஸ்ஃபானிஷ்க்கு எளிமையான வாக்கியங்களை மொழிபெயர்க்கும் ஆய்வு ஒழுங்குமுறையின் உருவாகத்திற்கும் (ரொசெட்டா 3) கவனகுவிப்பு செய்தது. அகராதிகள் சிறியாவை மற்றும் ஒழுங்குமுறை எல்லா சாத்தியமான மொழிபெயர்ப்புகளையும் உருவாக்கும். 1989-இல் தொடங்கப்பட்ட இரண்டாவது கட்டம் ரொசெட்டா 3-இன் அதிக வலுவான ஒழுங்குமுறை உருவாக்கத்திற்கு ஒதுக்கியது. இது உண்மையான பயன்பாடிற்கான மூலமுன்மாதிரி ஒழுங்குமுறையின் (ரொசெட்டா 4) உருவாக்கத்திற்கு வழிவகுத்தது. இதன் இறுதி நோக்கம் இலக்கு மொழிகளை அறியாத பயன்பாட்டாளர்களுக்கு ஒரு ஒழுங்குமுறையை உருவாக்குவதாகும். இது பகுப்பாய்வின் போது ஒருமொழிய ஊடாடும் பொருள்மயக்க நீக்கத்தை உட்படுத்தும்; மேலும் பின் திருத்தம் தேவைப்படாத வெளியீட்டை உருவாக்குவதாகும். எல்லா ஆய்வுகளும் கோட்பாடு மற்றும் மொழியியல் அடித்தளத்தில் கவனக்குவிப்பு செய்தன.

3.18.2. மாண்டேகு இலக்கணம் (Montague grammar)

மாண்டேகு இலக்கணத்தின் முக்கியப் பண்பு பொருள்கோள்களை அமைப்பு உறவுகளுடன் இணைப்பதாகும். மாண்டேகு இலக்கணம் ஒரு வெளிப்பாட்டின் பொருள் அதன் பாகங்களின் பொருளின் செய்பாடாகும் என்ற கூட்டமைவின் கொள்கைக்குக் (principle of compositionality) கட்டுப்படும். பாகங்கள் தொடரியலால் விளக்கப்படுவதால், தொடரியலுக்கும் பொருண்மையியலுக்கும் நெருங்கிய உறவு இருக்கும்.

மாண்டேகு இலக்கணம் 'அடிப்படை வெளிப்பாடுகள்'-இன் ஒரு குழுமத்தையும் தொடரியல் விதிகளின் ஒரு குழமத்தையும் குறிப்பிடும். அடிப்படை வெளிப்பாடுகள் பொருண்மைகொண்ட அலகுகளாகும்; மற்றும் விதிகள் எவ்வாறு அடிப்படை வெளிப்பாடுகளிலிருந்து பெரிய வெளிப்பாடுகளை (மற்றும் இறுதியாக வாக்கியங்களை) உருவாக்க இயலும் என்று விதிமுறை செய்கின்றது. விதிகள் கிழிருந்து மேலாகப் பயன்படுத்தப்படுகின்றன.

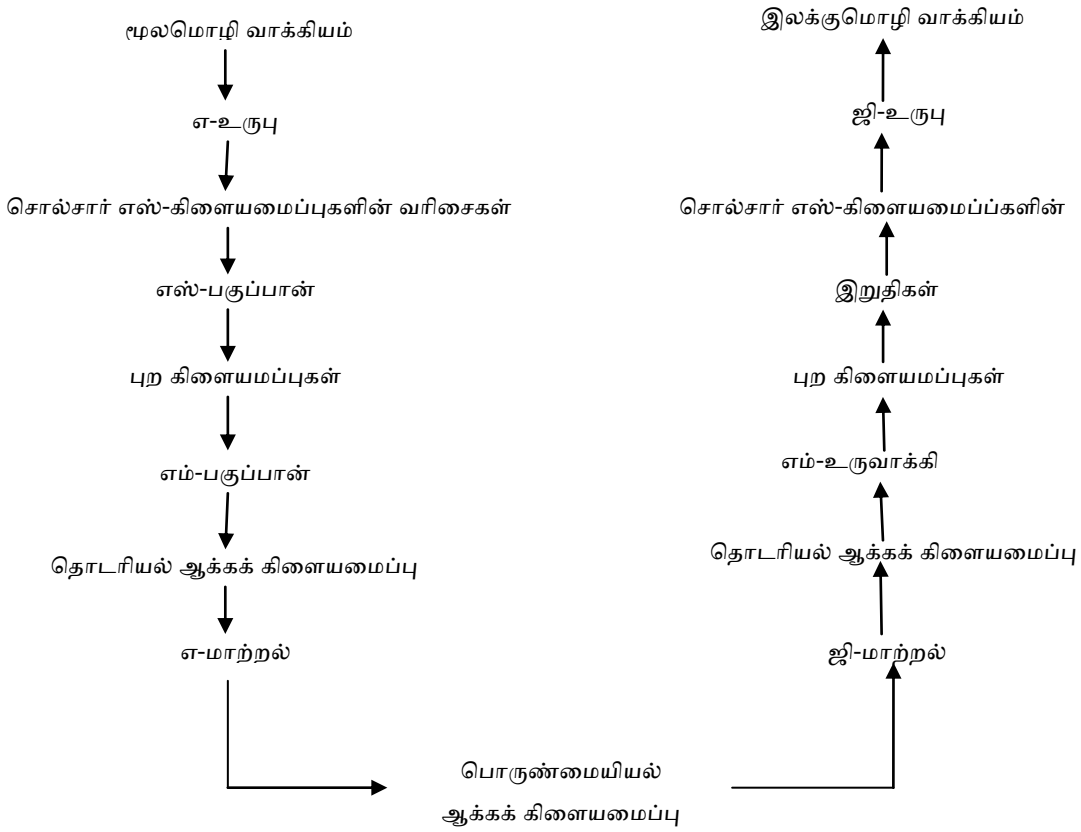
3.18.3. தலைகீழாக மாற்றல் மற்றும் ஒத்தவடிவுடமை (reversibility and isomorphism)

கூட்டமைவு மற்றும் இடைமொழியக் கொள்களைக் கொண்ட ரொசெட்டா இலக்கணங்கள் மூன்று பிற கொள்கைகளுக்கு இணங்கும். முதலாவது வெளிப்படை: மூலமொழி மற்றும் இலக்குமொழி இலக்கணங்கள் சுதந்திரமாக விளக்கப்படுவதுடன் மொழியியல் மற்றும் மொழிபெயர்ப்புச் செயற்பாங்குகள் இலக்கணங்களில் துல்லியமாக வெளிப்படுத்தப்பட்டுள்ளன. இரண்டாவது ஓர் இலக்கணக் கொள்கை: ஒரே இலக்கணம் வாக்கியங்களின் உருவாக்கத்திற்கும் பகுப்பாய்விற்கும் பயன்படுத்தப்படுகின்றது. அதாவது இலக்கணங்கள் தலைகீழாக மாற்றுவதற்கு உரியன. மிக முக்கியமான தேவை தொடரியல் விதிகளை தலைகீழாக மாற்றுதலாகும். மூன்றாவது கொள்கை ஒத்தவடிவுடமை. இது கூட்டமைவு கொள்கையைப் பின்பற்றும் தீர்மனத்திலிருந்து தொடர்கின்றது; மற்றும் பொருண்மை உருவாக்கக் கிளைகளை (semantic derivation trees) இடைமொழிய உருப்படுத்தங்களாகப் பயன்படுத்துவதைச் சாத்தியமாக்குகின்றது. இரண்டு வாக்கியங்கள் ஒரே பொருண்மை உருவாக்கக் கிளைகளைக் கொண்டிருந்தால் அவ்விரண்டும் ஒன்றுக்கொன்றான மொழிபெயர்ப்புகளாகக் கருதப்படும்; அதேபோன்று தொடர்புடைய தொடரியல் உருவாகக் கிளைகளும் (syntactic derivation trees) கருத்தப்படும்.

ரொசெட்டா ஆய்வாளர்களால் இடைமொழியம் (intelinguality) அல்லாது ஒத்தவடிவுடைமை தான் சட்டகத்தின் முதன்மை பண்பாக வலியுறுத்தப்பட்டது. இரண்டு வாக்கியங்கள் ஒன்றுக்கொன்றான மொழிபெயர்ப்பாக இருப்பதன் முக்கியமான கட்டுப்பாடு அவைகளுக்கு ஒத்தவடிவத் தொடரியல் உருவாக்கக் கிளை இருக்க வேண்டுமென்பதாகும்.

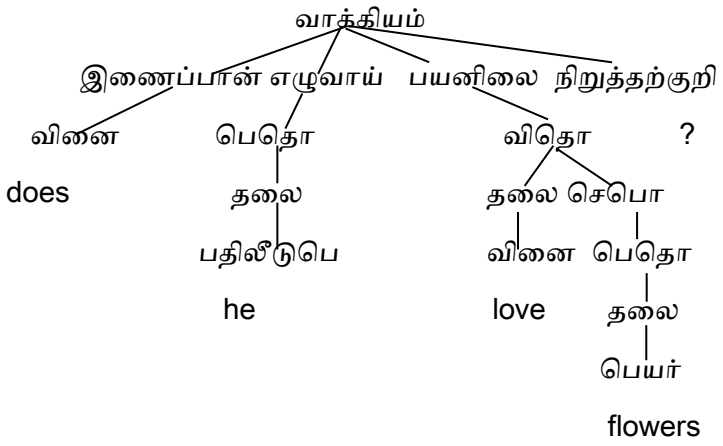
3.18.4. மொழிபெயர்ப்புச் செயற்பாங்குகள்

ரொசெட்டா மொழிபெயர்ப்பு செயற்பாங்கை பின்வருமாறு விளக்கலாம். மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு கீழ்வரும் படத்தில் எடுத்துக்காட்டியுள்ளது போல் எட்டு நிலைகள் உள்ளன.



முதல் நிலையில் உருபனியல் பகுப்பாய்வு (எ-உருபு), உள்ளீட்டுக் கோர்வைகள் என்பன 'சொல்சார் எஸ்-கிளையமைப்புகள்'-இன் (lexical S-trees) வரிசைகளை உருவாக்க பகுதிகளாகவும் (do, love, flower) ஒட்டுக்களாகவும் (-es, -s) சிதைக்கப்படும். மாண்டேகு இலக்கணத்தில் அடிப்படை வெளிப்பாடுகள் எளிய சொற்களாக இருந்தாலும், ரொசெட்டாவில் அவைகள் ஒன்று அல்லது அதற்கு மேற்பட்ட புலக்குறிப்புசெய்யப்பட்ட தனிமங்களைக்கொண்ட வரிப்படுத்தப்பட்ட கிளையமைப்புகளாக (எஸ்-கிளையமைப்புகள்/S-trees = surface trees) விளக்கப்படும். இரண்டாவது நிலை இரண்டு-நிலை தொடரியல் பகுப்பாய்வு ஆகும்; இதில் எஸ்-

பகுப்பான தற்காலிகப் பகுப்பாய்வின் குழுமங்களை உருவாக்கும். வகைப்பாட்டு ஒருசொல்போலி தீர்மானிக்கப்படுகின்றது (எ.கா. love ஒரு பெயரல்ல வினையாகும்); ஆனால் சொல்சார் அல்லது தொடரியல் பொருண்மை மயக்கம் தீர்மானிக்கப்படவில்லை.



தொடரியல் பகுப்பாய்வின் இரண்டாவது பகுதியில் எம்-பகுப்பான் தொடரியல் அடிப்படையிலான சரியான கிளையமைப்புகளைத் தேர்ந்தெடுக்கும்; வாக்கியத்தின் பொருளுளை உருப்படுத்தம் செய்ய அடித்தளத்தை ஏற்படுத்தும்; அதாவது அது ஒரு தொடரியல்சார் ஆக்கக் கிளையமைப்பை உருவாக்கும்.

தொடரியல் ஆக்கக் கிளையமைப்பு பொருண்மை உருப்படுத்தத்தை அதாவது ஒரு பொருண்மையியல் ஆக்கக் கிளையமைப்பைத் தீர்மானிப்பதற்கு பயன்படுத்தப்படும்; இது ஒவ்வொரு ('பொருண்மையுள்ள') விதியை இடைமொழியின் தொடர்புடைய 'பொருண்மை நடவடிக்கை'–உடன் பொருத்தியும் மூலமொழி அடிப்படை வெளிப்பாடுகளுக்கு இடைமொழிய அடிப்படை வெளிப்பாடுகளை பதிலீடு செய்தும் செய்யப்படும். அடுத்தக்கட்டத்தில் பகுப்பாய்வு மாற்றலால் (எ-மாற்றல்) இது நிகழ்த்தப்படும்.

இறுதி ஆக்கக் கட்டங்கள் இறுதித் தொகுதியால் சொல்சர் எஸ்-கிளையமைப்புகளின் வரிசையின் உருவாக்கம் ஆகும்.

3.18.5. **அமைப்புப் பொருத்தங்கள்**

ரோசெட்டா அமைப்புகளைப் பேணிவைக்காத்தால், அது இலக்கண வகைபாடுகளை மாற்ற இயலும் மற்றும் மூல அமைப்பிலிருந்து மிகவும் வேறுபட்ட இலக்கு அமைப்புகளை உருவாக்க இயலும்.

3.18.6. துணை இலக்கணங்கள்

ரோசெட்டாவில் இலக்கண வகைப்பாட்டு வேறுபாட்டின் சிக்கல்கள் எம்-இலக்கணங்களை (m-grammars) "வருவரைவு துணை இலக்கணங்கள்" (projection subgramars)ஆக ஆக்குவதன் வாயிலாகக் கையாளப்படும்; இது ஒவ்வொரு முதன்மையான வகைப்பட்டிற்கு (வினை, பெயர், முன்னுருபு, பெயரடை, வினையடை) ஒன்று என அமையும். ஒவ்வொன்றும் பல எண்ணிக்கையிலான துணை இலக்கணங்களைக் கொண்டிருக்கும்; எடுத்துக்காட்டாக, (வினைமுற்று கொண்ட) முழு எச்சத்தொடர்களை உருவாக்கவும், 'சிறிய (வினைமுற்று இல்லாத) எச்சத்தொடர்களிலிருந்து முழு வாக்கியங்களை உருவாக்கவும்.

3.18.7. விதி வகுப்புகள் (Rule classes)

எம்-இலக்கணங்களை துணை இலக்கணங்களாகப் பகுப்பதுடன் முன்னர் கூறியது போன்று 'அர்த்தமுள்ள' தொடரியல் விதிகளுக்கும் மாற்று விதிகளுக்கும் இடையில் வேறுபாடு காணப்படுகின்றது. பிந்தையது ஒரு குறிப்பிட்ட மொழிக்கான சிறப்பு விதிகளாகும்; அவை தொடரியல் செயல்பாடு மட்டும் செய்யும். முந்தைய ரோசெட்டா 2-இல் ஓரினச்சார்புக் கொள்கைகளின் (isomorphism principles) கண்டிப்பான பிரயோகம் பிற மொழிகளின் இலக்கணங்களில் அம்மாதிரியான விதிகளை உட்படுத்துவதை உள்ளடக்கும்; அங்கு அவை எந்த செயல்பாடையும் செய்வதில்லை. ஆனால் ரோசெட்டா 3-இல் அர்த்தமுள்ள மொழிபெயர்ப்புக்கு உகந்த தொடரியல் விதிகள் மட்டுமே ஓரினச்சார்புக் கட்டுப்பாடிற்கு ஆட்படுத்தப்பட்டுள்ளது. இந்த அர்த்தமுள்ள விதிகள் இணைதிறன் உறவுகள், நோக்கம், காலம், வினைப்பாடு போன்ற மொழியியல் நிகழ்வுகளின் வகைகளைக் கையாளும் விதிகளின் வகுப்புகளாக உருவாக்கப்படும். மேலும் அமைப்பு விதி எச்சத்தொடர்களுக்கு இடையில் பொருத்தங்களுக்கு மொழிபெயர்ப்பு உறவுகளைக் கட்டுப்படுத்துவதால் அறிமுகப்படுத்தப்படுகின்றது. ஒரே அர்த்தமுள்ள விதி வகுப்பைச் சார்ந்த வேறுபட்ட மொழிகளின் விதிகள் மட்டும் ஒன்றொடொன்று பொருந்தும்; எனவே ஒரே அர்த்தமுள்ள விதி வகுப்பைச் சாராத விதிகள் ஒன்றுக்கொன்றான மொழிபெயர்ப்பாக இருக்க இயலாது.

3.18.8. சொல்சார் இடமாற்றம் (Lexical transfer)

ஓரினச்சார்பு தேவை உட்கூறுகளான சொற்களுக்கு இடையில் பொருத்தம் இருக்கவியலாத மொழிபெயர்ப்பு நிகரன்களுக்கு எளிதில் கையாள இயலாத சிக்கல்களை முன்வைக்கத் தோன்றும். மிக வெளிப்படையான எடுத்துக்காட்டுகள் மரபுத்தொடர் வெளிப்படுகள் ஆகும்.

Spill the beans

Zijn mond voorbij praten

Lit.'one's mouth past talk' ('to talk past one's mouth'

இந்த வெளிப்பாடுகளின் நேரடியான அர்த்தங்களை அவற்றின் உறுப்புகளிலிருந்து இணைப்பாக்கமாக ஆக்க இயலாது என்பது சிக்கலாகும்; இருப்பினும் இந்த வெளிப்பாடுகள் முழுவதுமாக மொழிபெயர்க்கப்பட வேண்டும்.

ரோசெட்டாவில் மரபுத்தொடர்கள் இயல்பானவையாக உருப்படுத்தம் செய்யப்படவேண்டும்; அதே நேரத்தில் ஒரு தனியான 'அடிப்படை வெளிப்பாடாக' உருப்படுத்தம் செய்யப்படவேண்டும். இதன் விளைவாக ரோசெட்டா அகராதிகள் உள் அமைப்புகள் இல்லாத அடிப்படையான வெளிப்பாடுகளையும் (தொடக்கநிலை அர்த்தத்தையும்) உள் அமைப்புகளைக் கொண்ட கலவைத்தன்மையான அடிப்படை வெளிப்பாடுகளையும் கொண்டிருக்கும். ஒரு குறிப்பிட்ட வெளிப்பாடு நேரடியான அர்த்தத்தையும் மரபுத்தொடர் அர்த்தத்தையும் கொண்டிருந்தால் (எ.கா. kick the bucket) இலக்கணம் கூடுதலாக இணப்பாக்க அர்த்தத்தங்களையும் ஆக்கவேண்டும்.

இந்த அணுகுமுறை ஸ்பானிஷ் marugar போன்ற எளிய வினைகளுக்கும் get up early மற்றும் டச் vroeg opstaan என்ற தொடர்களுக்கும் இடையில் நிகரன்கள் போன்ற சொல்சார் இடமாற்றத்தின் சிக்கல்களைக் கையாள இயலும் என்று நம்பப்படுகின்றது. இந்த அணுகுமுறை ஓரனச்சார்பைத் (isomorphism) தக்கவைப்பதன் விருப்பத்தால் தெளிவாக ஊக்குவிக்கப்பட்டுள்ளது. இதே வழியில், பின்வரும் வாக்கியங்களின் மொழிபெயர்ப்பு நிகரன்கள் ஸ்பானிஷில் 'அருவத்தன்மையான முன்னுருபு' (abstract preposition) என்பதன் இருப்பிடத்தால் தற்காலிகமாகக் கையாளப்பட்டுள்ளது.

He ran across the square.

Hij rende het plein over

CruzÓ la plaza corriendo

ஸ்பானிஷின் ஆக்கத்தில் (ஸ்பானிஷுக்கு என்று குறிப்பாக அமையும்) இந்த முன்னுருபகளைக் கொண்டு இருக்கும் அமைப்புகளை இயக்கு வினையின் cruzar மற்றும் வினைப்பெயரைக் கொண்ட அமைப்பாக மாற்றும் மாற்று விதிகள் இருக்கவேண்டும்.

3.18.9. **கருத்துக்களும் முடிவுகளும்**

இறுதி இரண்டு பகுதிகளில் ஓரினசார்பு கொள்கையைப் பின்பற்றுவது ரோசெட்டாஆய்வாளர்களை ஒருசில 'உருக்குலைத்தல்களுக்கு' கொண்டுசென்றது என்பதைப் பற்றி பார்த்தோம். இது மொழிகளுக்குள்ளும் மொழிகளுக்கிடையேயும் ஒத்த ஆனால் ஒன்றல்லாத அமைபுகளைக் கையாள வேண்டி பகுதியான மற்றும் மேலுறல்செய்யும் துணைஇலக்கணங்களின் ஏற்பை உணர்த்தும். மற்றும் இது ஒருமொழிய அளவில் ஊக்குவிக்கப்பட இயலாத உறுப்புகள் அல்லது வாக்கியங்களின் உருவாக்கத்தை உணர்த்தும். ரோசெட்டா இலக்கணங்கள் மொழி நிகழ்வின் பெரிய பரப்பெல்லைகளைக் கையாளுவதற்கு விஸ்தரிக்கப்பட்டுள்ளது; மற்றும் முந்தைய ரோசெட்டா 2-இன் கட்டுப்பாடான ஓரினசார்பு திட்டம் நெகிழ்வான கருத்தாக்கத்திற்கு வழிவகுத்தது. முந்தையது மொழிகளுக்கிடையிலானது; அதாவது அவை பொருண்மை ஆக்கக் கிளைகளில் 'விதி அர்த்தங்கள்' என்பதுடன் பொருந்தும். பிந்தையது மொழிச்சிறப்பனது; அவை அண்மை உருப்படுத்தங்களில் ஒன்றுடனும் பொருந்தாது; மற்றும் பிறமொழிகளின் இலக்கணங்களின் விதிகளுடன் ஓரினச்சார்பு உடையதல்ல.

ரோசெட்டா மொழியிடை உருப்படுத்தம் ஒழுங்குமுறையின் மொழிகளின் ஓரினச்சார்பு இலக்கணங்களல் (isomorphic grammars) வரையறை விளக்கம் செய்யப்பட்டுள்ளது. மொழியிடை கூறுகளுக்கு வெளிப்படையாகவே உலகளாவிய தகுதி நிலை மறுக்கப்பட்டுள்ளது. ஜப்பானிய மற்றும் சீன மொழிகளின் ஓரினச்சார்பு இலக்கணங்களுக்கான பொருண்மையியல்சார் ஆக்கக் கிளை அமைப்புகள் டச் மற்றும் ஆங்கில மொழிகளின் ஓரினச்சார்பு இலக்கணங்கணங்களுக்கான பொருண்மைசார் ஆக்கக் கிளை அமைப்புகளிலிருந்து வேறுபடும் என்பதை எளிதில் எண்ணிப்பார்க்க இயலும். பல வழிகளில் ரோசெட்டாவின் திட்டம் மொழியியல் அடிப்படையிலான ஒழுங்குமுறையை ஒத்திருக்கும்: உருபனியல் ஆய்வு, புறத்தொடரமைப்புகள், அக அமைப்பு உறவு அடிப்படையிலான பொருண்மை ஆய்வு (இணைதிறன், காலம், வாய்ஸ், பயனிலைப் பங்கேற்பாளர் அமைப்பு போன்றன). தேவையான

முக்கியமான வேறுபாடு பிற ஒழுங்குமுறைகள் ஒற்றுமையல்லாத உருப்படிகளையும் அமைப்புகளையும் (சொல்சார் மற்றும் அமைப்புசார் இடமாற்றத்தால்) சமன்செய்யும் போது உருப்படிகளின் மற்றும் அமைப்புகளின் ஆக்க வரலாற்றை சமன்செய்கின்றது.

மொழியிடை அமைப்பொழுங்காக (intelingua system) ரோசெட்டா மூலமொழிப் பகுப்பாய்வும் இலக்கு மொழி உருவக்கமும் முற்றிலும் சுதந்திரமாக இருக்கவேண்டும் என்ற பொதுவான அனுமானத்திலிருந்து விலகிசெல்கின்றது. இலக்கணங்களும் வழிமுறைகளும் குறிப்பிட்ட மொழிகளின் மொழிபெயர்ப்பை நோக்கி வெளிப்படையாக ஆற்றுப்படுத்தப்படுவதை ஓரினச்சார்புக் கொள்கைவேண்டும். இக்காரணத்தால் ரோசெட்டாவை நேரடியான மொழிபெயர்ப்பு ஒழுங்குமுறையின் வகையாகக் மிகவும் சட்டபூர்வமாகக் கருத இயலும். வகைப்பாட்டுப் பொருந்தாமையைக் கையாளுதலும் மொழிபெயர்ப்பு மரபுத்தொடர்களின் பரிந்துரையும் இப்பண்பாக்கத்திற்கு ஆதரவு அளிக்கின்றது.

ஓரினச்சார்பின் வரையறை விளக்கம் ஒரு வெளிப்பாட்டிற்கு ஒன்றுக்கும் மேற்பட்ட 'அடிப்படை அர்த்தங்களின்' அல்லது ஒரு தொடரியல் செயல்பாட்டிற்கு ஒன்றிற்கும் மேற்பட்ட 'அர்த்த விதிகளின்' சாத்தியத்தையோ திறந்துவைப்பதாகத் தோன்றுகின்றது என்பதைக் காணலாம். இலக்கணம் ஜெர்மன் இலக்கணத்துடன் ஓரினச்சார்பாக இருக்க வேண்டுமானால் ஆங்கிலத்தில் wall என்பதற்கு இரண்டு அர்த்தங்கள் இருக்கவேண்டும். ஆனால் ஒரு பன்மொழிய ஒழுங்குமுறையில் இலக்கணம் பரிமாற்றப்படவேண்டுமானால் அர்த்தங்கள் பகுப்பாய்வின் போதும் உருவக்கத்தின் போதும் வேறுபடுத்தப்படவேண்டும்; மொழியிலிருந்து அல்லது மொழிக்கு மொழிபெயர்க்கும் போது வேறுபடுத்தப்படாத போதும் இது செய்யப்படவேண்டும். இது இடைமொழி அடிப்படையிலான எல்லா ஒழுங்கு முறைகளுக்கும் வழக்கமான சிக்கலாகும். இடமாற்ற அடிப்படையிலான விடை, ஓரினச்சார்புக் கொள்கை காரணமாகத் தள்ளுப்படி செய்யப்பட்டுள்ளது.

தலைகீழாக்கம் கட்டுப்பாட்டிற்கு கடினமான சிக்கல்களை எழுப்புகிறது. விதி பயன்பாட்டின் சில கட்டுப்பாடு துணை இலக்கண கட்டகமாக்கத்தால் (subgrammar modularization) அறிமுகப்படுத்துட்பட்டுள்ளது; இருப்பினும் எல்லா விதிகளும் விருப்பமாகதான் இருக்கவேண்டுமா என்ற கேள்வி எழுகின்றது. எம்-இலக்கணம் சுதந்திர உருவாக்க ஒழுங்குமுறைகள்; ஆனால் கணினி ஆற்றல் கட்டாய விதிகளை உட்படுத்துவதை ஆதரிக்கின்றது.

எதிர்பாராதவிதமாக உருவாக்கத்தின் போது கட்டாயமாக இருக்கவேண்டிய விதி பகுப்பாய்வின் போது விருப்பாக இருக்கவேண்டி உள்ளது; இதை தலைகீழ்மாற இலக்கணத்தால் கையாள இயலும் என்பதில் தெளிவு இல்லை. இருப்பினும் இது ரோசெட்டாவிற்கு என்று தனிப்பட்ட சிக்கல் அல்ல.

ரோசெட்டா ஒழுங்குமுறையை மதிப்பீடு செய்ய முயலுவதில் சில குறிப்பிட்ட சிக்கல்கள் உள்ளன. கருத்துரு நிலையில்/மட்டத்தில் சாத்தியமான அனுகூலங்களையும் பாதகங்களையும் புரிந்துகொள்வது கடினமாகும்; இதன் காரணம் புதுமையான கோட்பாடு சார் கொள்கைகளின் (Innovative theoretical principles) ஊடாட்டம் ஆகும். இக்கொள்கைகளில் பெரும்பான்மையானவை இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியில் முன்னர் பின்பற்றப்படவில்லை. இது வரை சிறிய செயல்முறை இலக்கணங்களே (demonstration grammars) உருவாக்கப்பட்டுள்ளன; அவற்றின் கணினிசார் நடைமுறைப் படுத்தம் (computational implementation) அறியப்படவில்லை. ரோசெட்டாவிலிருந்து கணினிசார் அம்சங்கள் பற்றிய சில உள் அறிக்கைகள் உள்ளன; இவை முக்கியமாகத் தலைகீழ் பகுப்பாய்விகள் மற்றும் உருவக்கிகளின் ஒழுங்குபடுத்தல் சம்பந்தப்பட்ட தாகும். வெளிப்படையான நிரல் மொழி புரோலாக இருக்கலாம் என்பது வெளிப்படையாக இருந்தாலும் எந்த நிரல் மொழி பயன்படுத்தப்பட்டுள்ளது என்பதில் கூட தெளிவில்லை.

செயல்முறை நடைமுறைபடுத்தத்தில் (practical implementation) ரோசெட்டா பயன்பாட்டாளர்களுடன் ஊடாட்ட முறையில் (interactively) செயல்பட விரும்புகின்றது. தற்போது ஊடாட்டம் பொருண்மை மயக்க நீக்கத்திற்கான துணைக்கருவியாகப் பகுப்பாய்வின் போதுதான் கருதப்படுகின்றது. ஒரு தனி விடுவரலுக்கு உருவாக்கத்தைக் கட்டுப்படுத்துவதன் நோக்கம் இல்லாதது போன்று தோன்றுகின்றது; அதாவது சில உள்ளார்ந்த குறிப்பிடப்பட்ட அளவுகோல் அடிப்படையில் மிகச்சிறந்த' மொழிபெயர்ப்பு.

ரோசெட்டா அணுகுமுறை பல இயந்திர மொழிபெயர்ப்பு ஆய்வாளர்களை ஊக்கப்படுத்தியுள்ளது;

3.19. டிஎல்டி (DLT)

3.19.1. வரலாற்றுப் பின்னணி

டிஎல்டி (DLT) என்பது டிஸ்ட்ரிப்யூட்டட் லாங்வேஜ் டிடான்ஸ்லேஷன் (Distributed Language Translation) என்பதன் சுருக்கம். இது அட்ரெஜ் மென்பொருள் நிறுவனம் (Utrecht softwre company) புரொ வூர் சிஸ்ட்மொண்டிவிக்கெலிங் (Buro voor Systemontwikkeling (BSO) என்பதன் ஆய்வுத்திட்டமாகும். எ.பி.எம். விட்கம்-இன் (A.P.M. Witkam) தொடக்க கட்ட ஆய்வு 1979-இல் தொடங்கப்பட்டது. 1985-இல் இவ்வாய்வுத் திட்டத்திற்கு நெதர்லாண்ட் பொருளாதார விவகாரங்களின் அமைச்சிலிருந்து ஆறு ஆண்டுகளுக்கான ஒப்பந்தம் கிடைத்தது. இதன் தொடக்கப் பணி 1987-இல் ஆங்கிலம்-பிரஞ்சு முன்மாதிரி ஒழுங்குமுறையையும் 1993-இல் ஒரு வியாபாரநோக்குசார் பதிப்புருவை உருவாக்குவதாகும். டிஎல்டி ஆய்வுத்திட்டத்தின் நீண்டகால நோக்கம் ஐரோப்பிய மொழிச்சமூகத்தை சார்ந்த மொழிகளுக்கு (பிரஞ்சு, ஜெர்மன், ஆங்கிலம், இத்தாலிய மொழி) இடையிலும் இறுதியில் பிற மொழிகளுக்கு இடையிலும் மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதாகும். புரோலாக் நிரல் மொழியில் எழுதப்பட்ட முன்மாதிரி கட்டுப்படுத்தப்பட்ட ஆங்கிலத்திலிருந்து (இலகுவாக்கப்பட்ட ஆங்கிலத்திலிருந்து) பிரஞ்சு மொழிக்கு மொழிபெயர்க்க வடிவமைக்கப்பட்டது. இது டிசம்பர் 1987-இல் செய்முறை விளக்கம் அளிக்கப்பட்டது. அந்நாளிலிருந்து டில்டி குழு முன்மாதிரி ஒழுங்குமுறையின் கொள்கை ஓரளவுக்கு வேறுபட்ட அடிப்படையில் வியாபாரம்சார் ஒழுங்குமுறையை உருவக்க வேண்டி செயல்பட்டுவருகின்றது.

டிஎல்டி தரவு கருத்துப்பரிமாற்ற வலைபின்னலில் (data communication networks) தனிப்பட்ட கணிப்பொறிகளில் பயன்படுத்துமாறு ஒரு ஊடாட்ட பன்மொழிய ஒழுங்குமுறையாக (interactive multilingual system) (அமைவதை நோக்கமாகக் கொண்டிருந்தது. இது ஒழு மொழிபெயரிப்புக் கருவியாக அல்லாமல் முதன்மையாக 'தகவல்' ஏடுகள் (சுருக்கங்கள், அறிக்கைகள், கையேடுகள்) அல்லது வியாபாரம் சார் செய்திகள் இவற்றின் இடைமொழி கருத்துப்பரிமாற்றத்தில் ஒருமொழியப் பயன்பாட்டாளர்களுக்குக் கருவியாக அமைவதை நோக்கமாகக் கொண்டிருந்தது. பகுப்பாய்வு மற்றும் உருவாக்கத்தின் செயற்பாங்குகள் வேறுபட்ட முனையங்களில் நடக்கின்றது என்ற நிலையில் மொழிபெயர்பு 'விநியோகிக்கப்பட்டது' (distributed) ஆகும். ஒருமொழியப் பயன்பாட்டாளர் ஒரு பனுவலை (எடுத்துக்காட்டாக

ஆங்கிலத்தில்) ஒரு முனையத்தில் உள்ளீடு செய்வார்; அங்கு அது எஸ்பெராந்தோ மொழி அடிப்படையிலான ஒரு இடைமொழியில் (intelingua) (அல்லது இடைப்பட்ட மொழியில்) உடனடியாக மொழிபெயர்க்கப்படும். பகுப்பாய்வும் மொழிபெயர்ப்பும் ஒரே நேரத்தில் (real time) நடைபெறும்; மொழிபெயர்ப்பு ஒழுங்குமுறை வாக்கியங்கள் முழுமையானதா அல்லவா என்பதைப் பார்க்காமல் பனுவலை அது உள்ளீடு செய்யப்படும் போதே மொழிபெயர்க்க முயற்சிக்கும். மொழிபெயர்ப்பு ஒழுங்குமுறை தீர்க்க இயலாத சிக்கல், இடைமொழியோ இலக்குமொழியோ எந்தமொழியில் பனுவல் மொழிபெயர்க்கப்படுகின்றது என்பதையோ அறியாத பயன்பாட்டாளர் ஊடாட்டத் தொடர்பு உரையாடல் (interative dialogue) வழி உள்ளீடுசெய்வதாகும். சில நேர்வுகளில் ஒழுமொழிய ஊடாட்டத் தொடர்புகள் மொழிபெயர்ப்புச் சிக்கல்களை எளிமையாக்குவதற்கும் நீக்குவதற்கும் வேண்டி மூலப் பனுவல்களை மாற்றி அமைப்பதற்குக் கொண்டு செல்லலாம். (இடைமொழியிலுள்ள) பனுவல் மற்றொரு முனையத்திற்கு கடத்தப்படகிறது; அங்கு மற்றொரு பயனர் இடைமொழியிலிருந்து தேவைப்படுகிற மொழிக்கு (எ.கா. பிரஞ்சுமொழிக்கு) மொழிபெயர்ப்பை தொடங்கிவைப்பார். பெறுபவரால் மொழிபெயர்ப்பு ஒழுங்குமுறையுடன் உரையாடல் செய்வது சாத்தியம் அல்ல; பின் திருத்தியமைப்பு எதிர்பார்க்கப்படுவதில்லை.

டிஎல்டி ஒரு கட்டக ஒழுங்கமைப்பாகும் (modular system); இதில் மூலமொழி மற்றும் இலக்கு மொழி கட்டகங்கள் (modules) பகுப்பாய்விற்கும் வேண்டி கொள்கை அடிப்படையில் உருவாக்க இயலும் மற்றும் இருக்கின்ற கட்டகங்களை பாதிக்காமல் இடைமொழியிலிருந்து உருவக்க இயலும். இந்த மொழிபெயர்ப்பு ஒழுங்குமுறையை வகைப்பாட்டியியல் அடிப்படையில் ஒற்றுமையுள்ள அல்லது ஒற்றுமையில்லாத பிற மொழிகளுக்கும் எளிதில் நீட்சி செய்வததை நோக்கமாகக் கொண்டுள்ளது. எந்த இலக்கு மொழிக்கும் தானியக்க முழு மொழிபெயர்ப்பைச் சாத்தியமாக்கும் படி தெளிவான மற்றும் பொருண்மை மயக்க இல்லாத ஒரு முழுமையாக வெளிபடுத்தும் இடைமொழியைத் தேவைகள் முற்கோள் செய்யும்.

3.19.2. இடைமொழி

பெரும்பாலான இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளில் இடைப்படா உருபடுத்தங்கள் உண்மையில் இடைமொழிசார்தது அல்ல. பொதுவாக அமைப்பு உருப்படுத்தம் மொழிச் சுதந்திரமானது. எ.கா. பயனிலை-பங்கேற்பாளர்

அமைப்பு; ஆனால் சொற்கள் அல்ல. பெரும்பாலான நேர்வுகளில் சொல் பெயர்ப்பு இருமொழிய அகராதிகள் அடிப்படையிலானது; (மொழிச் சுதந்திரமான) இடைமொழிய சொற்கள் இல்லை. முரணாக, டிஎல்டியில் எஸ்பெரந்தோ இடைமொழி அதற்குரிய சுதந்திரமான அமைப்புகள் மற்றும் சொற்களுடன் 'இயற்கை மொழி' போன்றது.

பெரும்பாலும் இடைமொழி பிற எந்த தனி மனித மொழிகளையும் விட வெளிப்படையாக இருக்கவேண்டும் என்று வாதிடப்படுவதுண்டு; அது பனுவல்களுக்குள் மற்றும் மொழி ஒழுங்குமுறைக்குள் (உட்படையான உறவுகளையும் உள்ளடக்கிய) எல்லா மொழியியல் சார் உறவுகளையும் (linguistic relationships) வெளிப்படுத்த இயலவேண்டும். இதன் அர்த்தம் உலகளாவிய அளவில் சரியான பொருண்மையியல் சார் மூலஅடிப்படைகளை (semantic primitives) உட்படுத்திய தருக்க உருப்படுத்தங்கள் (logical representations) என்பது அனுமானிக்கப்படுகின்றது. டிஎல்டி ஆய்வாளர்கள் இந்த அனுமானம் தவறாகும் என்று வாதிடுவர். மூல அடிப்படைகளாக பகுப்பாய்வு செய்வது நடைமுறையில் சாத்தியம் என்றாலும் இது 'ஊகிக்க இயலாத பெரிய அகராதிகளுக்கும் முடிவில்லாத ஆனால் பெரிய அளவிலான பயனற்ற செயற்பாங்குகளுக்கும்' கொண்டுசெல்லும். எந்த செயற்கை மொழியும் அதன் அடிப்படையாக அமையும் மற்றும் அது வரையறை விளக்கம் செய்யப்படும் மனித மொழிகளைவிட கூடுதல் வெளிப்படையானதாக இருக்க இயலாது என்று என்று அவர்கள் வாதிடுவர்; அது மனித மொழியின் திறனைத் தாண்டி அது செல்ல இயலாது. எனவே செயற்கை மொழியை எப்போது மனித மொழியில் மொழிபெயர்க்க இயலும் ஆனால் எப்போதும் மனித மொழியை முழுவதுமான செயற்கை மொழியாக மொழிபெயர்க்க இயலது என்று வாதிடுவர். இவ்வாறு டிஎல்டி ஆராய்ச்சியாளர்களின் கருத்தில் இடைமொழிக்கு 'மனித மொழியின்' பண்பு இருந்தால் தான் அது மனித மொழியைப் போல் வெளிப்படையாகவும் வெளியிடும் திறன் உள்ளதாகவும் இருக்க இயலும்.

டிஎல்டி அதன் இடைமொழியாக எஸ்பெராந்தோவைத் தெரிந்தெடுத்துள்ளது; அது 'இயற்கை' மொழியின் வெளிப்படுத்தும் திறனையும் (expressiveness) தேவைப்படும் முறைமையையும் (regularity) நிலையானதன்மையும் (consistency) ஒருங்கிணைந்ததாகும் என்று வாதிடுவர். இடைமொழியாக எஸ்பெராந்தோவின் அனுகூலங்களளாகப் பின்வருவனவற்றைக் கூறுவர்: (அ) ஒரு பகுதி 'இயற்கை' மொழியாக அதற்கு உருவாக்கப்பட்ட தருக்க

இடைமொழியை மிஞ்சும் அளவுக்கு வெளிப்படுத்தும் திறனும் வளமையும் நெகிழ்ச்சியும் உண்டு, (ஆ) பொதுவான இந்தோ-ஐரோப்பிய வேர்கள் அடிப்படையில் அது தயார்நிலையில் உள்ள தரப்படுத்தப்பட்ட சொற்றொகுதியைத் தருகின்றது (standardized vocabulary); (இ) அது சீரானது (regular) மற்றும் நிலையானது (consistent); (ஈ) அது தன்னியக்கமானதும் (autonomous) பிற மொழிகளிலிருந்து சுதந்திரமானதும் ஆகும்; (உ) அதைப் பிற மனித மொழிகளைப் போல் கற்கவும் புரியவும் இயலும்.

ஒரு இயந்திர மொழிபெயர்ப்பு இடைமொழியாக எஸ்பெராந்தோவிற்கு ஒருசில பிரதிகூலங்கள் (குறைகள்) உள்ளன. டில்டி தேவையான மாற்றங்களைச் செய்துள்ளது மற்றும் அதன் சொல்சார் அடிப்படையை விரிவாக்கியும் உள்ளது.

3.19.3. மொழிபெயர்ப்பு ஒழுங்குமுறையின் வடிவமைப்பு (System design)

இடைமொழி (மாற்றம் செய்யப்பட்ட எஸ்பராந்தோ) ஒரு அருவத்தன்மையான உருப்படுத்தமாக இல்லாமல் சீராக்கம் செய்யப்பட்ட மொழியாக இருப்பதால் அதன் விளைவாக மூலமொழிப் பனுவல்களின் பகுப்பாய்வும் இலக்குமொழிப் பனுவல்களின் உருவாக்கமும் இரண்டு 'மொழிபெயர்ப்பு ஒழுங்குமுறைகள்' ஆகும்: இலக்குமொழியிலிருந்து எஸ்பெராந்தோ மற்றும் எஸ்பெராந்தோவிலிருந்து இலக்கு மொழி. இதன் விளைவாக டிஎல்டி மொழிபெயர்ப்பு ஒழுங்குமுறையை அதன் மையத்தில் மாற்றப்பட்ட எஸ்பிராதோ கொண்ட இருமொழிய இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் வலைப்பின்னலாகக் (network of bilingual MT systems) கருதலாம்.

1988-வரை முன்மாதிரிக்காக (prototype) உருவாக்கப்பட்ட டிஎல்டி மொழிபெயர்ப்பு ஒழுங்குமுறை இந்த இரண்டு பாகங்களும் இருமொழிய மொழிபெயர்ப்பு ஒழுங்குமுறைகளிலிருந்து சுதந்திரமானதல்ல. 1988-இல் இருந்து ஆராச்சித் திட்டம் வேறுபட்ட (இருமொழிய அறிவு வங்கி) (Bilingual Knowledge Bank) கருத்தாக்கத்தில் வேலைசெய்யும்; இதில் இரண்டு பகுதிகளும் ஒரே வழியில் செயல்படும்.

இப்பகுதியில் உண்மையான எஸ்பெராந்தோ அடிப்படையிலான மாதிரி (Eperanto-based model) பற்றி விளக்கப்படும்.

முன்மாதிரி ஆங்கிலம்-பிரஞ்சு மொழிபெயர்ப்பு ஒழுங்குமுறையின் அடிப்படையான செயலாக்கக் கட்டங்கள் பின்வருவனவாகும்.

1. மூல மொழி இலக்கணப் பகுப்பாய்வு: ஒரு அதிகப்படுத்தப்பட்ட நிலைமாற்ற வலைப்பின்னலான (Augmented Transition Network (ATN)) இலக்கணப் பகுப்பான் ஆங்கிலச்சொற்கள், உருபனியல் சார் மற்றும் தொடரியல் சார் பண்புகூறுகள் இவற்றை அறிந்துகொள்ளும் மற்றும் (எழுவாய், செயப்படுபொருள், அடை போன்ற) சார்பு உறவுகளை (dependency relations) அடையாளம் காணும்; ஒரு சார்புக் கிளை அமைப்பை (dependency) உருவாக்கும்; தொடரியல் சார் பொருண்மை மயக்கம் இருந்தால் மாற்றுக்களை (alternatives) வழங்கும். பொருண்மையியல் உள்படுத்தப்படுவதில்லை; இது பொருண்மையியல் சத்தியம் (semantic plausibility) அல்லது புள்ளியியல் நிகழ்தகவு (statistical probability) இவற்றைப் பொருட்படுத்தாது எல்லா சாத்தியமான பகுப்பாய்வுகளையும் உருவாக்கும்.

2. ஒருமொழிய (மூலமொழி) கிளையமைப்பு மாற்றங்கள்: இந்த மட்டத்தில் ஒருமொழிய மாறிகள் (monolingual variants) பொதுவான வடிவங்களாகச் சுருங்கும்; எ.கா. *can not, cannot* மற்றும் *can't என்பன can not* ஆக மாறும்; துணைவினைக் கட்டமைப்புகள் (auxiliary verb constructions) புலக்குறிப்பு (labelled) செய்யப்பட்ட பண்புகூறுகள் கொண்ட ஒரு தனி வினையாகச் சுருங்கும்; எ.கா. *has been eaten* என்பது *eat* (முடிவுற்ற நிகழ்காலம், செயப்பாட்டு வினை) எனச் சுருங்கும்.

3. இருமொழிய கிளையமைப்பு மாற்றங்கள்: இது மெட்டாடாக்ஸோரின் பணியாகும்; இது ஆங்கிலச் சொற்களை எஸ்பெராந்தோ நிகரன்களால் இடம்பெயர்க்கும்; மற்றும் ஆங்கில தொடரியல் சார்புப் புலக்குறிப்புகளை (உறவுகளின் வெளிப்படையான காட்டிகளாக) எஸ்பெராந்தொவின் தொடரியல் சார்புப் புலக்குறிப்புகளால் (labels) இடம் பெயர்க்கும். இது கிளையமைப்பின் மறு ஒழுங்குகளையும் (rearrangements) செயற்பாட்டுச் சொற்களின் செருகலையும் தொகுகருத்துரை (entail) செய்யும். ஒவ்வொரு ஆங்கில கூறுக்கும் பெரும்பாலும் பல மொழிபெயர்ப்பு மாற்றுகள் (translation alternatives) இருப்பதன் காரணமாக (அதாவது ஆங்கிலத்தின் சொல்சார் பொருண்மை மயக்கம் காரணமாகவும் ஆங்கிலத்திலிருந்து எஸ்பெராந்தோவின் சொல்சார் மாற்றப் பொருண்மை மயக்கம் காரணமாகவும்) பல மாற்ற எஸ்பெராந்தோ கிளையமைப்புகள் (alternative Esperato trees) இருக்கும். இது முதல் மட்டமாக இருப்பதால் பொருண்மை சார் அல்லது பயன்வழியியல் சார் (pragmatic) தெரிவு செயல்படுத்தப்படவில்லை.

4. பொருண்மையியல்-பயன்வழியியல் சொல் தெரிவு: வழங்கப்பட்ட மாற்ற இடைமொழி கிளைகளிலிருந்து தரப்பட்ட சூழலில் அதிக சாத்தியமானது சொல்சார் அறிவு வங்கியில்

5. பொருண்மைமயக்கநீக்க உரையாடல்:

6. ஒருமொழிய (இடைமொழி) கிளையமைப்பு மாற்றங்கள்:

7. இடைமொழியின் கிளையமைப்பு நேராக்கம்: இடைமொழிக் கிளையமைப்பை எஸ்பெராந்தோவின் நேரான வடிவுக்கு மாற்றுதல் சரியான சொல் வரிசையைத் தீர்மானித்தல் மற்றும் புலக்குறிப்புகள் மற்றும் பண்புக்கூறுகளின் பட்டியலை நீக்குதல் இவற்றை உட்படுத்தும்.

8. செம்மை சரிபார்த்தல்: பாதுகாப்பிற்கு வேண்டி ஒவ்வொரு வக்கியம் அதன் தொடரியல் அடிப்படையிலான நல்லாக்கச் சரிபார்புக்காக ஒரு இலக்கணப் பகுப்பான் மூலம் செலுத்தப்படுதல்.

9. குறியாக்கம் மற்றும் வலைப்பின்னல் கடத்தல்/செலுத்தல்: ஏற்றுக்கொள்ளப்பட்ட வாக்கியங்கள் மின் கடத்தலுக்கு மாற்றப்படுதல்.

10. குறியத்திறவு: எஸ்பெராந்தோ பனுவல் மற்றொரு முனையத்தில் பெறப்படுதல்.

11. எஸ்பெராந்தோ இலக்கணப் பகுப்பான்: குறியத்திறவு செய்யப்பட்ட கோர்வை சார்பு கிளையமைப்பாக மாற்றப்படும்; ஏனென்றால் உள்ளீடு ஒப்பீட்டளவில் பொருண்மை மயக்கம் இல்லாததால் இந்தச் செயல்முறை விரைவானதாகும்.

12. ஒருமொழிய (இடைமொழி) கிளையமைப்பு மாற்றங்கள்: இந்த நிலை இரண்டாம் நிலையைப் பிரதிபலிக்கும்; தற்சமயம் எஸ்பெராந்தோவுக்கு.

13. இருமொழிய கிளையமைப்பு மாற்றங்கள்: மெட்டாடாக்சர் தனிப்பட்ட எஸ்பெராந்தோ கிளையமைப்பிலிருந்து மாற்று பிரஞ்சு சார்பு கிளையமைப்புகளின் ஒரு தொகுப்பை உருவாக்கும்; எ.கா. சொல்சார் இடமாற்றம் ஒரு எஸ்பெராந்தோ சொல்லுக்கு ஒன்றுக்கும் மேற்பட்ட சொற்களைக் குறிப்பிடும்.

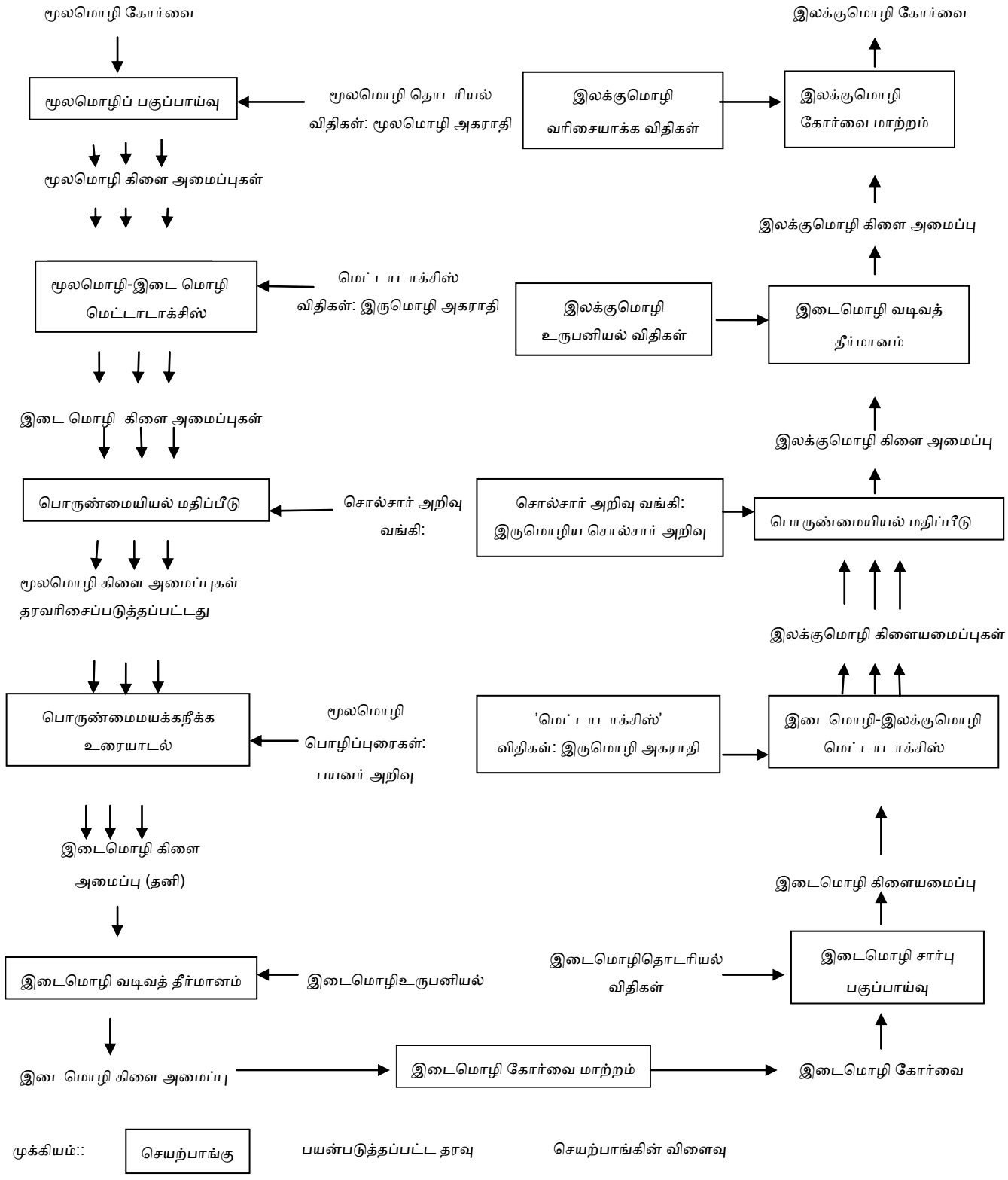
14. பொருண்மையியல்-பயன்வழியியல் சார் சொல் தெரிவு: இந்நிலையில் சொல்சார் அறிவு வங்கியில் உள்ள சொல் அமைப்பொழுங்குத் தகவலின் அடிப்படையில் சரியான பிரஞ்சு சொற்கள் தெரிவுசெய்யப்படும்.

15. ஒருமொழிய (இலக்குமொழி) கிளையமைப்பு மாற்றங்கள்: இது சரியான பிரஞ்சு சார்பு கிளயமைப்புகளின் சரிகட்டல் மற்றும் சொல் இயைபு மற்றும் உடன்பாட்டு உறவுகள் மற்றும் பண்புக்கூறுகளின் செருகல் இவற்றை உள்ளடக்கும்

16. இலக்கு மொழியின் கிளையமைப்பு நேராக்கம்: பிரஞ்சு கிளையமைப்பு தேவையான குறுக்கம் மற்றும் அசை கேடுகள் உடன் நேராக்கம் செய்யப்படும்.

இயந்திர மொழிபெயர்ப்பின் இரண்டு பகுதிகளும் பிரத்தேகமானவை/ வேறுபட்டவை ஆனால் சர்வசமான 'மொழிபெயர்ப்பு ஒழுங்குகுமுறையல்ல: மூலப் பனுவலை இடைமொழிப் பனுபலாக மாற்றுவது (கட்டம் 1இலிருந்து 8 வரை) இடைமொழிய பனுவலை இலக்குமொழிப் பனுவலாக மாற்றுவதும் (கட்டம் 11இலிருந்து 16 வரை) ஒற்றுமையானதல்ல. உருபனியல் மற்றும் தொடரியல் அம்சங்களைப் பொறுத்தவரையில் நெருங்கிய இணைகள் காணப்படுகின்றன: (அ) சார்புப் பகுப்பு (1இலிருந்து 11 வரை) (ஆ) மூலக் கிளையமைப்பு மாற்றல்கள் (2-ம் 15-ம்), (இ) இருமொழி கிளையமைப்பு மாற்றல்கள் (3ம் 13ம்), (உ) இலக்குக் கிளையமைப்பு மாற்றல்கள் (6ம் 15ம்), (எ) கிளையமைப்பு நேர்வரிசையாக்கம். கட்டங்களில் முக்கியமான வேற்றுமைகள் பொருண்மையியல்-பயன்வழியியல் விளக்கம் மற்றும் பொருண்மை மயக்க நீக்கம்; மூல மொழி மற்றும் இலக்கு மொழியில் எந்த செயற்பாங்குகளும் நிறைவேற்றப்படவில்லை; ஆனால் எல்லாம் எஸ்பெராந்தோ கெர்னல் ஒழுங்குமுறையில் நிறைவேற்றப்படும். முதல் பகுதியில் பொருண்மையியல் செயற்பாங்கின் கனமான சுமை இலக்குமொழியின் இறுதியில் நிறைவேற்றப்படும் (கட்டம் 4); இரண்டாவது பகுதியில் செயற்பாங்கு மூலமொழியின் இறுதியில் நடைபெறும். இதன் விளைவாக மூலமொழிக்கும் இலக்கு மொழிக்கும் குறிப்பிடத்தகுந்த தொகுதிகள் உருபனியல் மற்றும் தொடரியல் வடிவங்கள் மற்றும் அமைப்புகளின் மொழிச் சிறப்புக் கையாளுகைகளில் கவனக்குவிப்புச் செய்ய இயலும். எல்லா உள்ளடக்க (பொருண்மை) ஆய்வும் மாற்றமும் மொத்த ஒழுங்குமுறையின் இடைமொழி உட்கூறில் நடைபெறும்.

ஒவ்வொரு கட்டத்திலும் பயன்படுத்தப்படும் அடிப்படைச் செயற்பாங்குகள், தொகுதிகள் மற்றும் தரவுகள் என்பனவை திட்ட அடிப்படையில் கீழ் வரும் படத்தில் உள்ளது போல் எடுத்துக்காட்டப்பட்டுள்ளது.

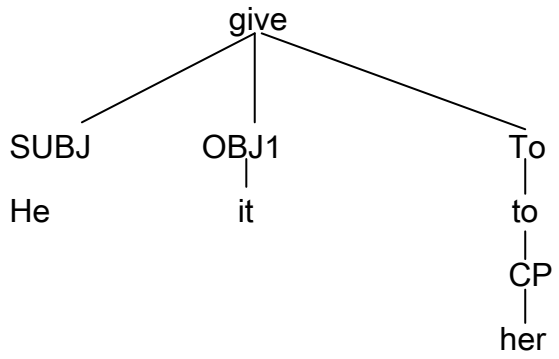

மூலமொழி கோர்வை
மூலமொழிப் பகுப்பாய்வு
மூலமொழி தொடரியல் விதிகள்: மூலமொழி அகராதி
மூலமொழி கிளை அமைப்புகள்
மூலமொழி-இடை மொழி மெட்டாடாக்சிஸ்
மெட்டாடாக்சிஸ் விதிகள்: இருமொழி அகராதி
இடை மொழி கிளை அமைப்புகள்
பொருண்மையியல் மதிப்பீடு
சொல்சார் அறிவு வங்கி:
மூலமொழி கிளை அமைப்புகள் தரவரிசைப்படுத்தப்பட்டது
பொருண்மைமயக்கநீக்க உரையாடல்
மூலமொழி பொழிப்புரைகள்: பயனர் அறிவு
இடைமொழி கிளை அமைப்பு (தனி)
இடைமொழி வடிவத் தீர்மானம்
இடைமொழிஉருபனியல்
இடைமொழி கிளை அமைப்பு
இடைமொழி கோர்வை மாற்றம்
இடைமொழி கோர்வை
இலக்குமொழி கோர்வை
இலக்குமொழி வரிசையாக்க விதிகள்
இலக்குமொழி கோர்வை மாற்றம்
இலக்குமொழி கிளை அமைப்பு
இலக்குமொழி உருபனியல் விதிகள்
இடைமொழி வடிவத் தீர்மானம்
இலக்குமொழி கிளை அமைப்பு
சொல்சார் அறிவு வங்கி: இருமொழிய சொல்சார் அறிவு
பொருண்மையியல் மதிப்பீடு
இலக்குமொழி கிளையமைப்புகள்
'மெட்டாடாக்சிஸ்' விதிகள்: இருமொழி அகராதி
இடைமொழி-இலக்குமொழி மெட்டாடாக்சிஸ்
இடைமொழி கிளையமைப்பு
இடைமொழிதொடரியல் விதிகள்
இடைமொழி சார்பு பகுப்பாய்வு
இடைமொழி கோர்வை
முக்கியம்::
செயற்பாங்கு
பயன்படுத்தப்பட்ட தரவு
செயற்பாங்கின் விளைவு

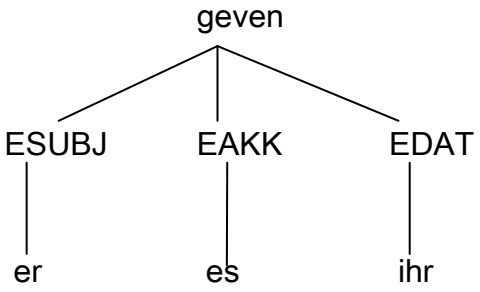
3.19.4. சார்பு பகுப்பாய்வு (dependency parsing)

டிஎல்டியில் எல்லா கட்டங்களிலும் அமைப்புகளின் உருப்படுத்ததிற்குப் பயன்படுத்தப்படும் வடிவவாதம் சார்பு தொடரியல் அடிப்படையிலானது. சார்புக் கிளையமைப்புகளைப் பயன்படுத்துவதற்கான அடிப்படை வாதங்கள்/பங்கெடுப்பாளர்கள் (arguments) அமைப்பு உறவுகளாகும் (எழுவாய், செயப்படுபொருள், அடை போன்றன); அவை உறுப்பு அமைப்புகளை (constituency structures) (பெயர்த்தொடர், முன்னுருபுதொடர்) விட அதிக அளவில் பகுப்பாய்விற்கும் மாற்றத்திற்கும் மையமாக உள்ளது. டிஎல்டியில் ஒவ்வொரு மொழியின் தொடரியலும் அர்த்தங்களைக் குறிப்பு செய்யாமல் (அதாவது சுத்தமாக வருகைமுறை அடைப்படையில்) அமைப்பொழுங்கில் பிற மொழிகளிலிருந்து சுதந்திரமாக உருவாக்கப்படுகின்றது. இவ்வாறு அண்மை உறவுள்ள மொழிகளில் உள்ள ஒரேமாதிரியான அமைப்புகளுக்குக்கூட அமைப்பு வேறுபாடுகளும் புலக்குறிப்புகளும் உள்ளன. பின்வரும் ஆங்கிலம் மற்றும் ஜெர்மன் மொழிகளின் உருப்படுத்தங்களை ஒப்பிடவும்:

1a. He give it to her.

2a. Er gibt es uhr.





டிஎல்டி பகுப்பான் எடிஎன் பகுப்பானின் ஒரு மாதிரியின் அடிப்படையிலானது. மூலக் கோர்வைகளின் இந்தப் பகுப்பாய்வில், பகுப்பான் மூலமொழியின் சொற்களின் 'தொடரியல் அகராதி' பதிவுகளில் அமைந்துள்ள தகவல் மீது இயங்குகிறது; பதிவுகள் திரிபுறாத வேர் வடிவுகளில் இருக்கும்; எனவே, பல இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளில் உள்ளது போல் உருபனியல் ஆய்வு, தொடரியல் (சார்பு) ஆய்வுக்கு முன்னர் செய்யப்படுகின்றது. பன்முகப் பகுப்பாய்வுகள் அமைப்புசார் நிலைமாற்றத்தின் அடுத்தநிலைக்குக் கடத்தப்படுகின்றன.

3.19.5.மெட்டாடாக்ஸிஸ் (Metataxis)

மெட்டாக்சிஸ் டிஎல்டியில் விதி ஒழுங்கமைப்புகளுக்குத் தரப்பட்பட்டுள்ள பெயராகும்; இது இரண்டு மொழிகளின் சார்பு தொடரியல்களை இனைக்கின்றது மற்றும் அமைப்புகளை மாற்றும் இயக்கமுறை 'மெட்டாடாக்ஸர்' என்று அழைக்கப்படுகின்றது. விதி ஒழுங்கமைப்புகள் ஒரு மொழி இணகளுக்கும் மற்றும் ஒரு மொழிபெயர்ப்புத் திசைக்கு குறிப்பிடப்பட்டது இவ்வாறு ஆங்கிலத்திலிருந்து பிரஞ்சு மூலமுன்மாதிரிக்கு இரண்டு மெட்டாடாக்ஸிஸ் விதி ஒழுங்கமைப்புகள் உருவக்கப்பட்டுள்ளன: ஆங்கிலந்திலிருந்து எஸ்பெராந்தோவுக்கு மற்றும் எஸ்பெராந்தோவிலிருந்து பிரஞ்சுக்கு.

மெட்டாடாக்ஸர் மூலமொழிப் பனுவல்களின் சார்புக் கிளையமைப்புகளை இலக்கு மொழிப் பனுவல்களின் சார்புக் கிளையமைப்புகளை சார்புக் கிளையமைப்புகளாக மாற்றுகின்றது. விதிகள் எல்லா மட்டங்களிலும் பயனுள்ளதாக உள்ளன: சொல், வாக்கியம் மற்றும் பனுவல். மெட்டாடாக்ஸிச் விதிகள் (அ) சொல்லின் தொடரியல் வகைப்பாட்டை மாற்றலாம் (ஆ)அதன் உருபனியல் வடிவை மாற்றலாம், (இ) அதன் தொடரியல் செயல்பாட்டை மற்றலாம், (ஈ) சார்பு உறவுகளின் கட்டமைப்பை (configuration) மாற்றலாம், (உ) சொற்களையோ உருப்படிகளையோ சேர்க்கவோ நீக்கவோ செய்யலாம் மற்றும் (ஊ) சார்பு கிளையமைப்புகளை

இணைக்கவோ பிரிக்கவோ செய்யலாம். மெடாடஸிஸ் மீதான கட்டுப்பாடுகள் மூல மற்றும் இலக்கு மொழிகளின் தொடரியல்களின் நல்லுருவாக்கக் கட்டுப்பாடுகளால் (well-formedness conditions) திணிக்கப்படுகின்றன.

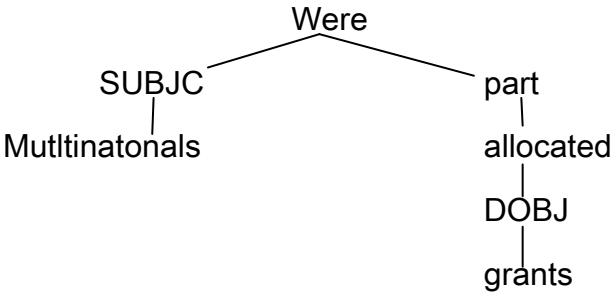
மெட்டாடாக்ஸிஸ் இரண்டு நெருக்கி ஒருங்கிணைக்கப்பட்ட செயற்பாங்குகளில் சொற்கள் மற்றும் கிளைகளின் மீது இயங்குகின்றன: இருமொழிய அகராதியிலிருந்து எடுக்கப்பட்ட ஒன்று அல்லது அதற்குக் கூடுதல் இலக்குமொழி நிகரன்களால் இடம்பெயர்த்தல் மற்றும் தொடரியல் அடிப்படையில் தற்காலிகமாகச் சரியான கிளைகளின் உருவாக்கம். வேறுபட்ட நிகரன்களின் விருப்பத்தேர்வுகள் கிளையமைப்பு அமைப்புகளில் மாற்றங்களை ஊக்குவிக்கின்றது மற்றும் சில கிளையப்பு மாற்றங்கள் சொற்களில் மாற்றங்களை வேண்டுகின்றது. சொற்கள் மீதான கட்டுப்பாடுகளும் கிளையமைப்புத் துண்டுகளாக முறைப்படுத்தப்படுகிறது. எடுத்துக்காட்டாக ஆங்கிலத்திற்கு பின்வருமாறு வரும்.

Sell---to---[ ]

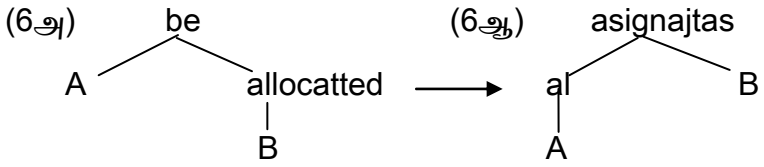
மெட்டாடாக்ஸிஸ் பொருண்மை மயக்கங்களைத் தீர்ப்பதுடன் அக்கறைகாட்டவில்லை: இருமொழிய அகராதி இரண்டுக்குக் கூடுதலான மொழிபெயர்ப்புக்களைக் கொண்டிருந்தால் மெட்டாடாஸிஸ் ஒன்றிற்கு மேற்பட்ட அமைப்பு உருப்படுத்தங்களை உருவாக்கும். மெட்டாடாக்ஸிஸின் நோக்கம் உருப்படுத்தங்கள் தொடரியல் அடிப்படையில் சரியாக இருக்கவேண்டும் என்பதை உறுதிசெயவதாகும். மாற்று உருப்படுத்தங்களுக்கு இடையிலான விருப்பத்தேர்வு SWESIL என்ற நிபுண ஒழுங்கமைப்பால் பொருண்மையியல் மற்றும் பயன்வழியியல் அடிப்படைகளின் மீது செய்யப்படுகின்றது.

ஆங்கிலத்திலிருந்து எஸ்பெராந்தோவுக்கு மெட்டாடாஸிஸை எடுத்துக்காட்ட 4ஆசார்பு பகுப்பாய்வு கொண்ட 4அவின் எஸ்பெராந்தோவாக மொழிபெயர்ப்பு 5ஐக் கருத்தில் கொள்ளவும்; இங்கு ஆங்கில எழுவாய் முன்னுருபு நிரப்பியாக (முன்னுருபு al ‘to’உடன்) மற்றும் ஆங்கில நேரடி செயப்படுபொருள் எழுவாயாகயாக மாற்றப்படுகின்றது. தீம்-ரீம் சொல் வரிசை பாதுகாக்கப் படலாம்; ஏனென்றால் subvencioj என்பதன் எழுவாய் வேற்றுமை அது எழுவாய் என்பதைச் சுட்டிக்காட்டுகின்றது.

4அ) Multinational were allocated grants.

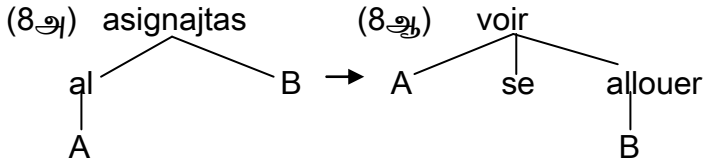


(5) Al mjltanciaj entreprenoj asignajtis subvencioj



எஸ்பிராந்தோ கிளையமைப்புகளை பிரஞ்சு கிளையமைப்புகளாக மாற்ற அதே போன்ற மெட்டாடாஸிஸ் விதி மூலமுன்மாதிரியின் இரண்டாம் பகுதியில் பிரயோகிக்கப்படும். எடுத்துக்காடாக, 4அ மற்றும் 5 இன் பிரஞ்சு நிகரன் 7ஐ உருவாக்க விதி 8 உள்ளதுபோல் இருக்கவேண்டும்.

(7) Les multinationals se sont vu allouer des subventions



அம்மாதிரியான விதிகள் ஒரு குறிப்பிட்ட அமைப்பு வகைகளுக்கும் சொற்களின் குறிபிட்ட கணங்களுக்கும் தெளிவாகத் திட்டவட்டமானதாகும். மேலும், குறிப்பிட்ட விதிகள் பயன்படுத்தப்படாவிட்டால் எல்லாம் மாற்றப்படவேண்டும் என்பதை உறுதி செய்ய வேண்டி பொது இயல்புநிலை விதிகளை மெட்டாடாக்ஸர் கொண்டிருக்கின்றது. ஒன்று மூல மொழி எழுவாய் சார்பு இலக்கு மொழி எழுவாய் சார்பாக மாறும்.

3.19.6.இடைமொழித் தரவும் SWESILஉம்

முதலில் கச்சிதமான இடைப்பட்ட உருப்படுத்தத்தின் வழியாகக் கருத்தப்பட்ட எஸ்பெராந்தோ திட்டம் உருவாக்கப்படுகையில் அதிக முக்கியத்துவம் அடைந்தது. ஒழுங்குமுறையின் இடைப்பட்ட கட்டங்களில் பொருண்மை விளக்கத்தில் கவனக்குவிப்பு செய்து ஒவ்வொரு கூடுதலான மொழிக்கும் அவை திரும்பசெய்யப்பட தேவைப்படாதிருக்க வேண்டி இடைமொழி ஒரு மொழியியல் 'அறிவு வங்கியை உருப்படுத்தம் செய்யும். மூலமுன்மாதிரி ஒழுங்குமுறையில் இரண்டு தரவுத்தளங்கள் இருந்தன: எஸ்பெராந்தோபனுவல்களின் LKB மற்றும் இருமொழிய எஸ்பெராந்தோ-பிரஞ்சு அகராதி. LKB எஸ்பெராந்தோ பனுவல்களின் 500,000 சொல் தரவுத்தொகுதியின் சார்பு ஆய்விலிருந்து பிரித்தெடுக்கப்பட்ட செயல்பாடு அல்லது உறவுபடுத்தியால் இணைக்கப்பட்ட பொருளடக்கச் சொற்களின் இணைகளைக் கொண்டது (எ.கா. கீழ்வரும் அட்டவணை). கணினிசார் எளிமை கருதி ஒவ்வொரு இணைகளின் நிகழ்வெண் பதிவுசெய்யப்படவில்லை. ஒரு எஸ்பெராந்தோ சொல் cambro ('room') சொல்லுக்கு ஒரு எடுத்துக்காட்டான எடுத்துக்காட்டு பின்வரும் அட்டவணையில் தரப்பட்டுள்ளது; இதில் a என்பது அடைமொழி உறவு படுத்தியாகும்.

அட்டவணை 1: word pairs for čambro

čambro a apada	‘adjoining room’
čambro a bela	‘beautiful room’
čambro a granda	‘large room’
čambro a hela	‘light room’
čambro a komforta	‘comfotable room’
čambro a nuda	‘bare room’
etc	

இருமொழிய அகராதியில் எஸ்பெராந்தோ மற்றும் பிரஞ்சு சொற்களின் இணைப்புகள் சொல்சார் மாற்றத்தின் போது செய்யவேண்டிய விருப்பத்தேர்வுகளில் சூழல்களைச் சுட்டிக்காட்டும் பனுவல் துப்புகளால் தொடரப்பட்டுள்ளன. இந்த துப்புகளும் எஸ்பெராந்தோவில் தரப்பட்டுள்ளன.

அட்டவணை 2: இருமொழிய இணைகளின் சூழல் துப்புகள்

akara a (doloro, malvarmo, vortoj, …)	→	vif
akara a (nazo, oreloj, turo)	→	pointu
akara a (spico, pipro, brando)	→	fort
akara a (dispute, batalo, krizo, …)	→	violens
akara a (ironio)	→	mordant

அட்டவணை எஸ்பெராந்தோ akara என்பது doloro ‘’pain’, malvarmo ‘cold’, vortoj ‘words’ என்பதன் அடையாய் வந்தால் vif என மொழிபெயர்க்கப்படவேண்டும்; nazo ‘nose’, oreloj ‘ears’, turo ‘tower’ என்பதன் அடையாய் வந்தால் pointu என மொழிபெயர்க்கப்படவேண்டும் spico ‘spice’, pipro ‘pepper’, brando ‘brandy’ என்பதன் அடையாய் வந்தால் fort என மொழிபெயர்க்கப்படவேண்டும் என்று காட்டுகின்றது.

இந்த இரண்டு தரவுத்தளங்களும் பொருண்மை செயற்பாங்கின் இரு கட்டங்களில் ‘நிபுண ஒழுங்குமுறை’ SWESILக்கு மூலங்களாகும்: ஆங்கிலத்திலிருந்து இடைமொழிக்கு மொழிபெயர்க்கும் போது மற்றும் இடைமொழிய உருப்படுத்தத்தைப் பிரஞ்சுமொழிக்கு மொழிபெயர்க்கும் போது.

3.19.6.1.**ஆங்கிலத்திலிருந்து எஸ்பெராந்தோவுக்கு பொருண்மையியல் செயற்பாங்கு (English to Esperanto semantic processing)**

அட்டவணை 1 போன்ற அட்டவணைகள் எஸ்பெராந்தோ சொல் உறவுகளின் ஏற்றுக்கொள்கையைப் பரிசோதிக்க மூலமுன்மாதிரியில் (prototype) பிரயோகிக்கப்படுகின்றது. மூலமுன்மாதிரியின் வடிவமைப்பின் தொடக்கக் கட்டத்தில் சொற்களின் படிநிலை அமைப்புகளுக்குக் குறிப்புரையால் கையாளப்படுகின்றது.

மூலமொழியின் (ஆங்கிலம்) எல்லா பொருண்மை மயக்கங்களும் (word sense disambiguation) SWESILஆல் தீர்க்க இயலாது; எனவே மூலமுன்மாதிரி 'உரையாடல் பொருண்மை மயக்கநீக்கம்' (dialogue word sense disambiguation) (கட்டம் 5) என்ற கட்டத்தை உட்படுத்தும்.

3.19.6.2.**எஸ்பெராந்தோவிலிருந்து பிரஞ்சுக்கு பொருண்மையியல் செயற்பாங்கு (Esperanto to French semantic processing)**

முதல் அரைப் பகுதியில் (மூலமொழியிலிருந்து இடைமொழிக்கு மொழிபெயர்ப்பு, கட்டம் 4) மொழிபெயர்ப்பில் பொருண்மை மயக்கநீக்கம் இடைமொழி LKBயில் பொருந்தும் சொல் இணையின் ஒருமொழிய அறிவை (monolingual knowledge) மட்டும் உட்படுத்தும்; இரண்டாவது அரைப் பகுதியில் (இடைமொழியிலிருந்து இலக்கு மொழிக்கு, கட்டம் 13) பொருண்மை மயக்கநீக்கம் இருமொழிய அறிவை (bilingual knowledge) உட்படுத்தும் மற்றும் வேறுபட்ட பொருத்தும் செயல்முறைகள் பிரயோகிக்கப்படும். இலக்கு மொழியில் (பிரஞ்சு) சொல் விருப்பத்தேர்வு பிரஞ்சு சொற்களின் ஒருமொழிய சாத்தியமான இணைப்புகளால் அல்லாமல் உள்ளீட்டு (எஸ்பெராந்தோ) சூழல்களுடன் சூழல் துப்புகளைப் பொருத்துவதால் தீர்மானிக்கப்படும்.

3.20. பிற வேறு ஒழுங்குமுறைகளும் ஆய்வின் திசைகளும்

இங்கு கூறப்பட்டுள்ள செய்திகளை இயந்திர் மொழிபெயர்ப்பு ஆய்வுச் செயல்பாடுகளின் சமகால விளக்கம் என்று கருதலாம் (Hutchins and Somers 1992: 314-333).

3.20.1. செயற்கை அறிவும் சிஎம்யு-இல்அறிவு அடிப்படையிலான இயந்திர மொழிபெயர்பு

மொழிபெயர்ப்பு மேம்பாட்டை உற்றுநோக்கும் பெரும்பாலோனவர்களுக்கு இயந்திர மொழிபெயர்ப்பின் தரத்தை மேம்படுத்த சாத்தியமான தொழிநுட்ப மூலவளம், செயற்கை

அறிவின் (Artificial Intelligence (AI)) சூழலில் இயற்கை மொழி ஆய்வு ஆராய்சி ஆகும். இயந்திர மொழிதொடர்பான திட்டங்களில் செயற்கை அறிவு ஆய்வாளர்களின் ஈடுபாடு 1970இன் தொடக்கத்தில் தொடங்கியது. ஸ்டான்போர்ட் பல்கலைக்கழகத்தில் யொரிக் வில்க்ஸ் என்பரின் செயல்பாடும் யேல் பல்கலைகழகத்திலும் ரோகெர் ஷாங் மற்றும் அவரது கூட்டாளிகளின் ஆய்வும் இதற்கு எடுத்துக்காட்டுகளாகும். அல்பாக் அறிக்கை இயந்திர மொழிபெயர்ப்பு அணுகுமுறைகளின் போதாமையை சுட்டிக்காடிய பின்னர் இது நிகழ்ந்தது. அக்காலத்தில் பலரால் அறியப்பட்ட முக்கியமான குறைபாடு 'பொருண்மைத் தடை' என்பதன் முக்கியத்துவம் ஆகும்.

செயற்கை அறிவு அணுகுமுறைகளின் அடிப்படையிலான நியாயப்படுத்தம் என்னவென்றால் மொழிபெயர்ப்பு ஒரு மொழியிலிருந்து மற்றொரு மொழிக்குப் பனுவலின் பொருளடக்கத்தை அல்லது பொருண்மையைத் தெரிவிப்பதுடன் தொடர்புள்ளபடியால் எந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையும் 1960களில் பார்-ஹிலெல், யங்வே மற்றும் பலர் வாதிட்டது போல் பனுவல்களின் பொருண்மைகளைப் புரிந்து கொள்ள வேண்டும். புரிந்துகொள்ளுதல் இல்லாமல் எந்த சாத்தியமான இலக்குமொழி வெளிப்பாடுகள் மூல பனுவலின் பொருண்மையுடன் மிக அண்மையில் பொருந்தும் என்பதை எந்த ஒழுங்குமுறையும் தீர்மானிப்பதை எதிர்பார்க்க இயலாது. செயற்கை அறிவு அணுகுமுறைகளின் சிறப்புப்பண்பு என்னவென்றால் பொருண்மையியல் சார்ந்த பகுப்பாய்வை பின்பற்றுவதாகும்: அறிவு அடிப்படைகளுக்குக் குறிப்புரையால் பனுவல்களின் பொருள்கோள் மற்றும் பனுவல்களின் 'பொருண்மையின்' மொழிச் சுதந்திரமான உருப்படுத்தங்கள்

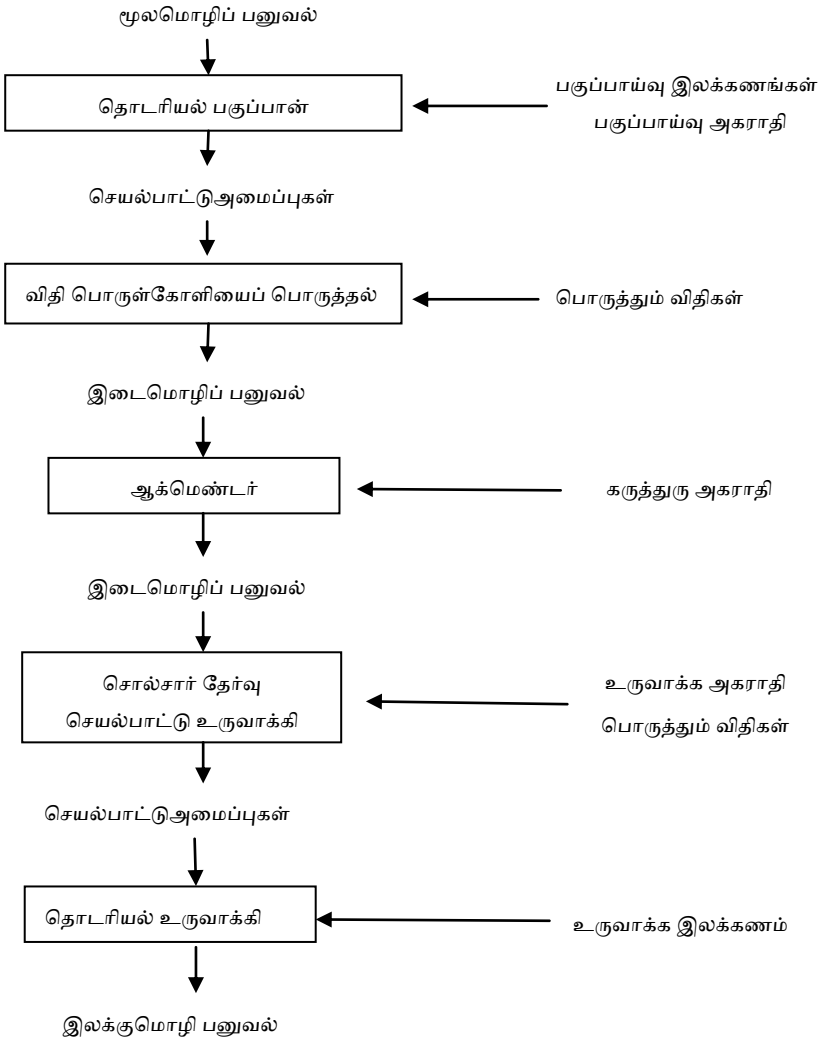
1980கள் மொழிபெயர்ப்புக்கு செயற்கை அறிவு அணுகுமுறையைகள் மீதான ஆய்வுகளில் தொடர்ந்த மற்றும் அதிகரிக்கும் செயல்பாட்டைக் கண்டது. ஐரோப்பாவிலும் (சில யுரொட்ரா திட்டத்துடன் தொர்புள்ள) ஜப்பானிலும் (குறிப்பாக மின்னணு-தொழில்நுட்ப ஆய்வுக்கூடம் (Electro-Technical Laboratory)) மற்றும் குறிப்பாக வட அமெரிக்காவிலும் இது நிகழ்ந்தது. பெரும்பாலான இந்த செயற்கை அறிவு ஊக்கப்படுத்தப்பட்ட ஆய்வுகள் சிறிதளவில் தான் இருந்தது; ஆனால் சில ஆண்டுகள் முக்கியமான மையம் பிட்ஸ்பர்கில் உள்ள கார்னி-மெலொன் பல்கலைக்கழகம் (Carnegie-Mellon University (CMU) ஆக இருந்தது.

தொடக்கத்தில் 1983இல் கோகேட் பல்கலைக்கழகத்தில் (Colgate University) தொடங்கப்பட்ட செயல்பாடு, கார்போனெல் மற்றும் ஸேர்கெய் நிரென்பர்க் (Carbonell and

Sergei) என்பவர்களின் கீழ் சிஎம்யு இயந்திர மொழிபெயர்ப்பு மையத்தில் (CMU Centre for Machine Translation) ஆய்வு தொடர்ந்து நிகழ்த்தப்பட்டது. 'இடைமொழி சட்டகத்தில் பொருண்மை சார்ந்த இயந்திர மொழிபெயர்ப்பு' ('machine-oriented MT in an intelingua paradigm') என்று விளக்கப்பட்ட நெறிமுறை அடிப்படையில் பரிசோதனை ஒழுங்குமுறைகள் அமைந்தன. உருவாக்கப்பட்ட ஒழுங்குமுறைகள் எடுத்துக்காட்டான இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் தோராயமாகக் கருதப்பட்டது. ஒழுங்குமுறையின் சில பாகங்கள் ஒத்தறி அடிப்படையில் முழுமையானவை; கள அறிவு மற்றும் அகராதிகள் மற்றும் மொழியியல்சார் ஆய்வின் பல இடங்கள் இவற்றை உள்ளடக்கிய பிற பகுதிகள் பரிசோதனை நிலையில் உள்ளன.

ஆங்கிலம் மற்றும் ஜப்பானிய மொழிகளுக்கு இரண்டு திசைகளிலும் செயல்படுகின்ற மூலமுன்மாதிரி தனிகணினி கையேடுகளின் மொழிபெயர்ப்புக்காக வடிவமைக்கப்பட்டது. இதில் 1500 கருத்துருக்களின் ஒரு சிறிய 'கள மாதிரி'யும் இருமொழிகளுக்கும் இருபது சொற்கள் உள்ள பகுப்பாய்வு மற்றும் உருவாக்க அகராதிகள் உள்ளன. ஒழுங்குமுறைகள் CommonLispஇல் எழுத்தப்பட்டுள்ளது; இலக்கண வடிவாதம், சொல்சார் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG))அடிப்படையிலானது. பொருண்மையியல் கட்டுப்பாடுகள் கொண்ட தொடரியல் பகுப்பாய்வி (syntactic parser), பொருண்மையியல் பொருத்தம் காட்டி (semantic mapper), மீதமுள்ள பொருண்மை மயக்கங்களுக்கு ஊடாடும் ஆக்மெண்டர் (interactive 'augmentor' for remaining ambiguities), சொல்சார் தேர்வு கொண்ட தொடரியல் அமைப்புகளை உருவாக்கும் ஒரு பொருண்மையியல் உருவாக்கி (semantic generator producing syntactic structures with lexical selection) மற்றும் இலக்குமொழி கோர்வைகளை உருவாக்கும் தொடரியல் உருவாக்கி (syntactic generator for producing target strings) என்பன அடிப்படைத் தொகுதிகள் ஆகும் (கீழ்காணும் படம்). கருத்துரு அகராதி மற்றும் பகுப்பாய்வு மற்றும் உருவாக்கம் அகராதிகளிலுள்ள பொருண்மையியல் தகவல்கள் மொழி சுதந்திரமானவை ஆனால் தலைப்பு களத்திற்குச் சிறப்பானவை. செயல்பாட்டு அமைப்புகளை (f-structures) இடைமொழிப் பனுவல்களாக மாற்றும் பொருத்தல் (mapping ruels) மொழி மற்றும் பொருண்மைக்களம் சுதந்திரமானவை. கருத்துரு அகராதிகளை உருவாக்கவும் ('அறிவு ஈட்டக் கருவி' ONTOS)

இலக்கணகளைத் தொகுக்கவும் தொகுதிகளையும் உட்கூறுகளையும் பரிசோதிக்கவும் CMU ஒழுங்குமுறை மென்பொருளால் ஆதரிக்கப்பட்டுள்ளது.



பகுப்பாய்வி இரண்டு உட்கூறுகளைக் கொண்டிருக்கும்: தொடரியல் பகுப்பாய்வி (syntactic parser) மற்றும் பொருண்மையியல் பொருள்கோளி (semantic interpreter), 'விதி பொருள்கோளிப் பொருத்துதல்' ('mapping rule interpreter). தொடரியல் பகுப்பாய்வி LFG-வகை இலக்கணத்தைப் பயன்படுத்துகின்றது மற்றும் LFG-வகை 'செயல்பாட்டு அமைப்பை' (f-structure)

உருவாக்குகிறது. எடுத்துக்காட்டாக வாக்கியம் (1), வாக்கியம் (2)இல் உள்ளது போன்று உருப்படுத்தம் செய்யப்பட்டுள்ளது.

(1) Remove the diskette from the drive.

(2) ((OBJ ((CASE ACC) (REF DEFINITE)
(DET ((ROOT THE) (REF DEFINITE)))
(ROOT DISKETTE) (PERSON 3) (NUMBER SINGULAR)
(COUNT YES) (PROPER NO))
(PPADJUNCT ((PREP FROM) (REF DEFINITE)
(DET ((ROOT THE) (REF DEFINITE)
(ROOT DRIVE) (PERSON 3) (NUMBER SINGULAR)
(COUNT YES) (PROPER NO)))
(VALENCY TRANS)
(MOOD IMPERATIVE)
(TENSE PRESENT)
(FROM INF)
(COMP-TYPE NO)
(ROOT REMOVE)

பொருண்மையியல் பொருள்கோளி பொருண்மை மயக்கதிற்கு வேண்டி எந்தச் செயல்பாட்டு அமைப்பு உட்கூறையும் பரிசோதனைசெய்கிறது மற்றும் மூலமொழிச் சொல்லுக்கும் கட்டமைப்புகளும் இடைமொழி அலகுகளை இடம்பெயர்க்கின்றது. எ.கா. புற எழுவாய்-பயனிலை அமைப்புகள் வேற்றுமைச் சட்டங்களால் இடம்பெயர்க்கப்படுகிறது ('செயலி', 'மையக்கருத்து', 'அனுபவிப்பவர்', போன்றன).

ஒழுங்குமுறையின் மைய நடுப்பகுதி இடைமொழிய பனுவல்களின் உருப்படுத்தம் ஆகும். இவை மூலப்பனுவல்களில் அறிவிக்கப்பட்டுள்ளது போன்று 'உண்மை நிகழ்வுகளின்' உருப்படுத்தம் ஆகும். அவை கருத்துரையின் (proposition) வலைப்பின்னல்களின் வடிவில் உள்ளன; அதாவது நிகழ்வுகள் அல்லது நிலைகள் அவற்றின் பங்கெடுப்பாளர் மற்றும் காரணம்சார், காலம்சார், இடம்சார் போன்றவற்றுடன் பிற நிகழ்வுகள் அல்லது நிலைகளுடன்

இணைக்கப்பட்டுள்ளன. உருப்படுத்தங்கள் 'கருத்துரு அகராதி'யிலிருந்து கருத்துருக்களின் எடுத்துக்காட்டுகளாக (நிகழ்வுகள், தனிநபர்கள், போன்றன) உருவாக்கப்படுகின்றது. பிந்தையது பொருண்மைக்களத்தில் (ஒழுங்குமுறையின் எழுவாய்ப் புலம்) நிகழ்வுகள் (events) மற்றும் இருப்புகள் (entities) பற்றிய அறிவின் தரவுதளம் ஆகும். நிலையான அறிவுத் தரவுத்தளத்திற்கும் (குறிப்பிட்ட பனுவல்கள் சாராத உறவுகளின் வலைப்பின்னல்) இயக்க இடைமொழி பனுவல் உருப்படுத்தங்களுக்கும் இடையில் தெளிவான வேறுபாடு உள்ளது. உள்ளீட்டுப் பனுவல்களின் தேவையான புரிதலை உறுதிசெய்ய தேவைப்படுகின்ற அறிவு கருத்துரைசார் அறிவைத் தாண்டி செல்லவேண்டும்; அது பயன்வழி மற்றும் கருத்தாடல் பொருண்மையை உட்படுத்தவேண்டும்; அதாவது வெளியிடப்பட்ட கருத்துரைகளுக்குப் (propositions) பேசுபவர்கள் மற்றும் கேட்பவர்களின் உளப்பாங்குகள், பேச்சு செயல்கள் (speech acts), கருத்து அமைப்புகள் (thematic structures) மற்றும் தனித்தனியான கூற்றுகள் ஒத்திசைவான பனுவல்களாக இணைக்கப்படும் வழிகள் இவற்றை உட்படுத்தும்.

இடைமொழி பனுவல் ஸ்லாட் மதிப்புகளுடன் சட்டங்களின் நேரியல் அல்லாத வலைப்பின்னலாக உருப்படுத்தம் செய்யப்படுகின்றது. பொதுவாக மதிப்புகள் கருத்துரு அகராதியில் பட்டியலிடப்பட்டுள்ளவைகளுடன் ஒத்திருக்கும். எ.கா. பனுவல்சார் rose பட்டியலிலிருந்து ஒன்றாக 'colour' மதிப்பு (வெள்ளை சிவப்பு மஞ்சள் நீலம் ...) கொண்டிருக்கும் 'flower' என்ற கருத்துருவுடன் இணைக்கப்பட்டிருக்கும் (இதில் rose என்பது ஒரு நேர்வாக அடையாளம் காணப்படும்). உள்ளீடு செய்யப்பட்ட பனுவல்களிலிருந்து கருத்துரை அல்லாத தகவலின் உருப்படுத்தம் சிஎம்யு திட்டத்தின் புதுமையான பண்புக்கூறாக கருதப்படும். அதே அறிவு அமைப்பு கருத்துரு அகராதியிலும் மற்றும் இடைமொழி பனுவல்களில் கருத்துரைத் தகவல்களாகப் பயன்படுத்தப்படுகின்றது. இவ்வாறு அகராதி, பேச்சுச் செயல்களின் ('அறிக்கை', 'வரையறை விளக்கம்', வேண்டல்-தகவல்', 'வேண்டல்-செயல்') சாத்தியமான மதிப்புகளை அடையாளங்காட்டும்; இதில் ஒன்று பனுவலில் உள்ள குறிப்பிட்ட வாக்கியத்திற்குப் பொருத்தமானதாகும். இதுபோன்று அகராதி பனுவல் இயைபின் குறியீடுகளுக்கு மதிப்புகளின் கணங்களைத் தருகின்றது: 'விரிவாக்கம்', 'ஒப்புமை', 'பொதுமை', 'முரண்பாடு', 'வழிவிலகிய போக்கு' போன்றன.

செயல்பாட்டு அமைப்பில் (f-structure) மீது பொருத்தும் விதி பிரயோகத்தின் விளைவு (2) தேர்ந்தெடுக்கப்பட்ட இடைமொழிப் பனுவல் (3) ஆகும்.

(3) [*REMOVE
(THEME (*DISKETTE (NUMBER SINGULAR)
(REFERENCE DEFINITE)))
(SOURCE (*DISKETTE-DRIVE (NUMBER SINGULAR)
(REFERENCE DEFINITE)))
(TENSE PRESENT)
(MOOD IMPERATIVE)]

எடுத்துக்காட்டுகள் காட்டுவது போன்று, அறிவு அடிப்படை 'drive' என்பதன் அடையளம் 'diskette drive' என்பதன் தலைப்புப் பொருண்மைக்களத்தைக் குறிப்பிடும் என்பதைச் செயல்படுத்துகின்றது.

ஆகுமெண்டரின் செயல்பாடு, உருவாக்கியின் உள்ளீட்டுக்கு ஒரு தனிப் பொருண்மை மயக்கமில்லாத இடைமொழிப் பனுவலை உருவாக்குவதாகும். பகுப்பாய்வியிலிருந்து வெளியீடு ஓரளவுக்கு மூலமொழியின் தொடரியல் கட்டுமானங்களைப் பிரதிபலிப்பதால் அது ஆர்குமெண்டரால் மொழி சுதந்திரமான வடிவில் மறுவடிவாக்கம் செய்யப்படவேண்டும். இரண்டாவதாக, ஆகுமெண்டர் இடைமொழியப் பனுவல்களின் நேர்மையான பொருண்மை மயக்கப் பகுதியைப் பொருண்மை மயக்கநீக்கம் செய்யவேண்டும். எடுத்துக்காட்டாக, tape என்பது இந்தப் பொருண்மைக் களத்தில் 'ஒட்டும் டேப்'ஐயோ 'காந்த டேப்'ஐயோகுறிப்பிடும் மற்றும் பகுப்பாய்வியின் பொருண்மையியல் செயற்பாங்கு வாக்கியம் (4) போன்றவற்றின் பொருண்மை மயக்கத்தை தீர்க்க இயலாது.

(4) Remove the tape from the diskette drive.

இங்குதான் கருத்துரு அகராதியில் பொதிந்துள்ள கருப்பொருளின் 'அறிவு' வேண்டப்படும். இதுவரை சிளம்யு திட்டம் ஆக்மெண்டரின் பொருண்மை மயக்கச் செயல்பாடுகளின் ஒரு பகுதியை மட்டுமே தானியக்கப்படுத்தப்பட்டுள்ளது; அதாவது வாக்கியங்களைக் கடந்த மற்றுப்பெயர் முற்சுட்டுக்களின் குறிப்பீடுகளை அடையாளம் காணத் தானியக்கப்படுத்தப்பட்டுள்ளது; இதன்

பலனாக தற்போது பெரும்பாலான ஆர்குமெண்டெட்டரின் செயல்பாடுகள் பயனர்களால் ஊடாட்டமாக நிறைவேற்றப்படுகின்றது.

உருவாக்கி சொற்களின் தேர்வாலும் பகுப்பாய்வியில் உள்ள பெருத்த விதிகளுச் (mapping rules) சமமான விதிகளின் பயன்பாட்டாலும் இலக்குமொழியின் செயல்பாட்டு அமைப்பை (f-structure) முதலில் உருவக்குகின்றது; பின்னர் புற அமைப்பையும் வெளியீட்டு பனுவலையும் உருவாக்குகின்றது. பகுப்பாய்வின் விளைவு சாத்தியமான பொருள்கோள்களின் (interpretation) பன்முக வெளியீடாக (multiple output) இருக்கையில், உருவாக்கம் ஒரு சரியான கோர்வை உருவாக்கப்பட்டதும் (அதாவது முதலாவது உருவாக்கப்பட்டதும்; அது மிக நல்லதாகவோ முழு உள்ளீட்டுப் பனுவலை வெளிப்படுத்தும் ஒன்றாகவோ இருக்கத் தேவையில்லை) நிற்கிறது. பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் ஒரே இலக்கணங்கள் தான் பயன்படுத்தப்படுகின்றன; ஆனால் ஜப்பானிய மொழிக்கு மட்டும் சிறிது விதிப் பிரயோகத்தின் பகுதி முன்பின்மாற்றம் (reversibility) இருக்கின்றது.

சிஎம்யு-இன் முக்கியத்துவம், கருத்துருசார் ('பொருண்மை') இடைமொழியில் உருப்படுத்தங்கள் வழி இயந்திர மொழிபெயர்ப்பின் ஆய்விலும் ஒரு பொருண்மைக் களத்தின் திட்டவட்டத்திலும் ஆனால் குறிப்பிட்ட மொழி சாராமையிலும் இருக்கின்றது. மனித மொழிபெயர்ப்பாளர்கள் அவர்கள் என்ன மொழிபெயர்க்கிறார்கள் என்பதை முழுவது புரிந்து கொள்ளத் தேவையில்லை என்று சுட்டிக்காட்டப்படுகின்றது. இரண்டாவதான ஆட்சேபனை செயற்கை அறிவு-வகை பொருண்மை நோக்கிய ஒழுங்குமுறைகள் மொழிபெயர்ப்புக்குப் பதிலாகச் சுருக்கவுரையைத் தான் உருவாக்கும் என்பதாகும். முக்கியமான ஆட்சேபனை மொழிச் சுதந்திரமான அறிவு அடிப்படைகளின் கட்டுமானங்கள் மிகக் கட்டுப்படுத்தப்பட்ட பொருண்மைக் களத்தைத் தவிர பிறவற்றுக்குச் சாத்தியமானதா என்பதாகும்.

3.20.2. பிஎஸ்ஓ-வில் எடுத்துகாட்டு அடிப்படையிலான மொழிபெயர்ப்பு

எடுத்துக்காட்டு அடிப்படையிலான நெறிமுறைகள் அறிவு-அடிப்படையிலான அணுகுறைக்கு மாற்றாகவோ மரபு விதி அடிப்படையிலான நெறிமுறைகளுக்குத் துணையாகவோ அமைய இயலும். ஜப்பான்சார் எடிஆர் திட்டத்தின் அங்கத்தினர் உட்பட பல்வேறு ஆய்வாளர்கள் அவற்றின் திறனை ஆய்கின்றனர். இதை எடுத்துக்காட்ட 'இருமொழிய அறிவு வங்கி'ஐ (Bilingual

Knowledge Base (BKB)) மொழிபெயர்ப்புக் கருவியாகப் பயன்படுத்தும் டிஎல்டி ஆய்வுத்திட்டத்திலிருந்து ஒரு முன்மொழிபை விளக்குவோம் (Hutchins and Somers 1992: 317).

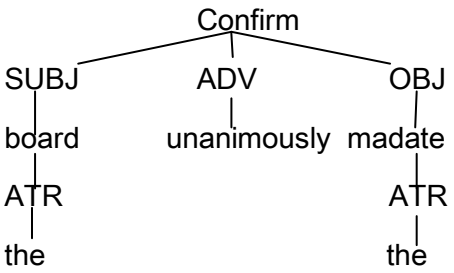
இருமொழிய அறிவு வங்கியின் நோக்கம் மொழிபெயர்ப்பு செயல்முறையில் எல்லா தொகுதிகளுக்கும் மொழியியல்சார் 'அறிவை'வின் முதன்மை மூலவளமாகச் செயல்படுவதாகும். இது அமைப்பு அடிப்படையில் (ஒரே போன்ற பகுப்பானால்) 'மொழிபெயர்ப்பு அலகு'களாகப் ('translation units) பகுப்பாய்வு செய்யப்பட்ட மற்றும் ஒன்றுக்கொன்று வரிசைப்படுத்தப்பட்ட இரு மொழிகளில் நிகரான பனுவல்களின் (equivalent texts) தரவுத்தொகுதியைக் கொண்டிருக்கும். மொழிபெயர்ப்பு அலகுகள், ஒரு மொழிபெயர்ப்பாளர் நிகரான மற்றும் ஒன்றுக்கொன்று இடம்பெயர்க்ககூடியன என்று கருதும் இரு மொழிகளின் பனுவல்களின் துண்டுகளாகும்; அவை அமைப்பு அடிப்படையில் சுதந்திரமானவை. இம்மாதிரியான வரிசைப்படுத்தப்பட்ட 'இருமப்பனுவல்கள்' (bitexts) முன்னர் தொழில்முறை மொழிபெயர்ப்பாளர்களுக்குத் துணைக்கருவியாக முன்மொழியப்பட்டது; அவைகளுக்கு மொழிபெயர்ப்பின் முந்தைய நடைமுறைகளை அணுகும் வசதி தரப்பட்டுள்ளது. டிஎல்டி ஆய்வாளர்கள், மூலமொழிப் பொருண்மை மயக்கங்களின் தானியக்கத் தீர்ப்புக்கும் சொல்சார் மற்றும் அமைப்புசார் மாற்றதிற்கும் (lexical and structurl transfer) இலக்குமொழி தேர்வுச் (target language selection) சிக்கல்களுக்கும் அமைப்பாக்கம் செய்யப்பட்ட இருமொழியத் தரவுத்தொகுதியைப் (bilingual corpus) பயன்படுத்த விரும்புகின்றனர்.

வரிசைப்படுத்துவதற்கு எடுத்துக்காட்டாக, ஆங்கில-பிரஞ்சு இருமொழிய வங்கியிலுள்ள வாக்கியங்கள் (5அ)உம் (5ஆ)உம் சார்புப் பகுப்பானால் (6அ)உம் (6ஆ)ஆக பகுப்பாய்வு செய்யப்படவேண்டும்.

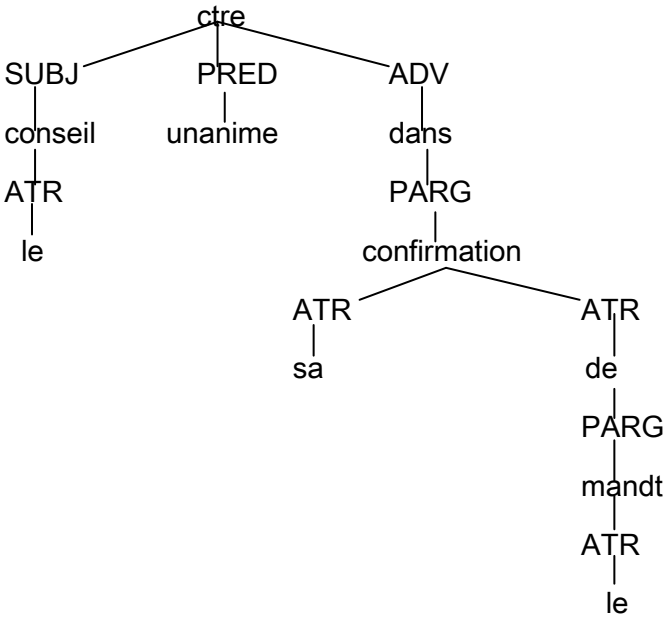
(5அ) The board unanimously confirms the madate.

(5 ஆ) Le conseil est unanime dans sa confirmation du mandat.

6அ)



(6ஆ)



அமைப்புக்கு வரிசைப்படுத்தப்பட்ட மொழிபெயர்ப்பு அலகுகள் இருமொழிய அமைப்புசார் நிகரன்களுடன் தொர்புபடுத்தும் (7) விதிகளால் ஆக்கப்படும்.

(7) the board ↔ le conseil

the le

unanimously confirm ↔ être unanime dans confirmation de

unanimously ↔ unanime

the mandate ↔ le mandat

முழு ஒப்பீட்டு ஆய்வின் மற்றும் பெரிய தரவுத்தொகுதியில் மொழிபெயர்ப்பு அலகுகளின் வரிசப்படுத்தலின் விளைவாகவும், (8)இல் உள்ள தரவு have…effect on… என்ற வடிவம் ஆங்கில வெளிப்பாடுகளுக்கு கிடைப்பதாக இருக்கும்.

(8) have a direct effect on ↔ *ont une influence directe à*
have a direct effect on ↔ *interéssent directement*
had a direct effect on ↔ *ont eu une repércussion direte sur*
has had a marked effect on ↔ *a largement influencé*
had a positive effect on ↔ *s est avérée positive dans*
had a highly negative effect on ↔ *en auraient été gravement affectés*
will have a direct effect on ↔ *influencera de façon déterminante*

மூலமொழியிலிருந்து இலக்குமொழி மாற்றத்தின் கட்டங்கள் முந்தைய டிஎல்ட் மாதிரி போல இருக்கும். வேறுபாடு ஒவ்வொரு கட்டத்திலும் பயன்படுத்தக்கூடிய தகவலின் வகையில் உள்ளது. தொகுதிகள் கிளையமைப்புகளை வரிசையாகக் கடந்துசெல்லாமல் பொதுவான தரவு அமைப்பின் மீது இயங்குகிறது. பல்வேறான தரவுத்தளங்கள் (அகராதிகள் மற்றும் பிற அறிவு அடிப்படைகள்) இருமொழிய அறிவு வங்கியில் ஒருங்கிணைக்கப்பட்டுள்ளது. நான்கு இயக்கநுட்பங்கள் ஊடாட்டமாக இயங்குகிறது: பகுப்பான், பனுவல் வல்லுனர், மெட்டாடாக்ஸர் மற்றும் பரிசோதிப்பவர். பகுப்பானும் மெட்டாடாக்ஸரும் தொடரியல் பகுப்பாய்வுக்கும் மாற்றத்திற்கும் முந்தைய பகுதிகளில் விளக்கப்பட்ட இயக்கநுட்பங்களுடன் கிட்டத்தட்ட பொருந்தும்.

வாக்கியங்களைக் கடந்த குறிப்பீடுசார் உறவுகளைக் கையாள வேண்டி முன்மொழியப்பட்ட பனுவல் நிபுணர் ஒரு புதிய தொகுதி ஆகும். எடுத்துக்காட்டாக, வாக்கியத் தொடர்ச்சி (9) தரப்படுகையில் பனுவல் நிபுணர் it என்ற மாற்றுபெயரின் சாத்தியமான முற்சுட்டுகளைத் தேடும்: இந்நேர்வில், பொருண்மையியல் இணக்கத்தன்மை இன்மை காரணமாக translation அல்லது screen; translator அல்ல.

(9) The translator may see the translation on his screen. It will be displayed in a special format.

பிரஞ்சு மொழியையோ ஜெர்மன் மொழியையோ மொழிபெயர்க்கும் போது ஒரு தீர்மானம் தேவை ஆகும்: il அல்லது er screenஐக் (*écran* அல்லது *Bildschirm*), குறிப்பிடும்; elle அல்லது sie *translation*ஐக் (traduction அல்லது *Ubersetzung*) குறிப்பிடும்.

பரிசோதிப்பவரின் செயல்பாடு, பகுப்பானிலிருந்து மிகச்சரியான பகுப்பாய்வுகளை தேர்ந்தெடுப்பதும் மெட்டாடாக்ஸர்-இலிருந்து மிகச்சரியான மாற்றங்களைத் தெரிந்தெடுப்பதும் மற்றும் பனுவல் நிபுணரிலிருந்து பனுவல்-இலக்கண நிச்சயமற்றவைகளைத் தீர்ப்பதும் ஆகும். இது இருமொழிய அறிவு வங்கிக் குறிப்பீட்டால் பொருள்கோள்கள் மற்றும் முன்மொழியப்பட்ட மொழிபெயர்ப்புகளின் பொருண்மையியல் மற்றும் பயன்வழியியல் சாத்தியங்களை மதிப்பீடு செய்கின்றது. இவ்வாறு ஒரு குறிப்பிட்ட நேர்வில் *around* என்பது *autour de* அல்லது *as vers*, *sharp as aigu*, *pointu*, *vif* அல்லது *aigre* என மொழிபெயக்கப்படவேண்டுமா என்பதை பரிசோதிக்க, பரிசோதகர் மூலமொழிப் பனுவல்களிலோ இலக்கு மொழி பனுவல்களிலோ (அல்லது இரண்டிலுமோ) ஒத்த சூழல்களின் எடுத்துக்காட்டுகளுக்கு வேண்டி தேடவேண்டும். செயல்முறை, *big* மற்றும் *large* (பிரஞ்சு மொழியின் மொழிபெயர்ப்பாக *grand* என்பதையும் ஜெர்மன் மொழியின் மொழிபெயர்ப்பாக *gross* என்பதையும்) என்பவைகளுக்கு இடையிலுள்ள விருப்பத்தேர்வையும் பிரஞ்சு மொழியின் *rapide* என்பதை மொழிபெயர்க்கையில் *fast*, *swift*, *rapid*, *quick* என்பனவற்றிற்கு இடயிலுள்ள நுட்பமான வேறுபாடுகளையும் பொதுவாக 'நடையியல்' என்று பண்பாக்கம் செய்யப்படும் சொல்சார் மாற்றங்களின் சிக்கல்களை கையாள இயல்வேண்டும். செயல்முறை சாத்தியமாக உட்படையான உறவுகளை அடையாளங் காண்கின்றது; எடுத்துக்காட்டாக, radiation protection என்பதில் கதிர்வீச்சுக்கு எதிராக ஏதாவது காக்கப்படவேண்டும்; ஆனால் wildlife protection என்பதில் எதற்கோ யாருக்காவது எதிராக வனவிலங்குகள் காக்கப்படவேண்டும். தொடர்புடைய பனுவல்கள் பின்வரும் தொடர்களை உட்படுத்தவேண்டும்: *steps taken to protect the inhabitants against radiation hazards* மற்றும் *reservations for the protection of wild birds and animals* (மற்றும் பறவைகள், விலங்குகள் இவற்றை வனவிலங்குகள் என்பதுடன் இணைப்பது). இந்த அணுகுமுறை (9)இல் எடுத்துக்காட்டப்பட்ட வாக்கியங்களுக்கு இடையிலான சிக்கல்களைக் கையாளும். எடுத்துக்காட்டாக *translation* என்பதையும் *text* என்பதையும் மற்றும் *the text was displayed*... போன்ற துண்டுகளையும் இணைக்கும் எடுத்துக்காட்டுகளைக் கண்டுபிடிப்பதன் மூலம். இந்த

எடுத்துக்காட்டுகள் மூலமொழியிலுள்ள பனுவல்களிலிருதோ இலக்குமொழியிலுள்ள பனுவல்களிலிருந்தோ வர இயலும் என்பது வலியுறுத்தப்படவேண்டும்; இருமொழியத் தரவுத்தொகுதி முழுவதுமாக மொழிகடந்த அறிவின் தரவுத்தளத்தையும் கருத்தப்பட்ட இரண்டு மொழிகளைப் பற்றிய மொழி அறிவின் மூலத்தையும் உருப்படுத்தம் செய்யவேண்டும்.

இருமொழிய அறிவு வங்கி அணுகுமுறையின் நன்மைகளாகப் பின்வருவன கருதப்படுகின்றன: (அ) ஒழுங்குமுறை கொள்கை அடிப்படையில் முன்பின்னாக மாற்ற (reversible) இயலும்; இரண்டு மொழிபெயர்ப்பு திசைகளிலும் ஒரே செயல்முறைகளும் பனுவல்சார் தகவல்களும் பயன்படுத்தப்படுகின்றன; (ஆ) எந்த அகராதிகளும் எந்த அறிவுத் தளங்களும் (knowledge bases) நிபுணர்களின் முயற்சியின் அதிக செலவீனத்துடன் தொகுக்கப்படவேண்டும்; முன்பின்முரண்பாடு, போதாமை இவற்றின் ஆபத்து தவிர்க்கப்படவேண்டும்; (இ) பகுப்பாய்வு விருப்பத்திற்குத் தகுந்தாற்போல் ஆழமாகவோ ஆழமின்றியோ இருக்கவியலும்; அனுபவம் அடிப்படையில் மாற்ற இயலும்; முழு தரவுத்தளம் கொண்டிருப்பதால் தகவல் ஏதும் இழப்பதில்லை; (ஈ) பனுவல் தரவுத்தொகுதிகளை (இதன் காரணமாகச் சொல்சார் பரப்பெல்லையை) குறிப்பிட்ட பயனர்களின் தேவைக்குத் தகுந்தாற்போல் தெரிந்தெடுக்க இயலும்; (உ) ஒப்புரு மற்றும் பல்பொருள் ஒருமொழியம் பற்றிய தீர்மானங்கள் மிதமிஞ்சியது; பனுவல் தரவுத்தொகுதியின் 'துணைநிலை மொழிகள்' ('sublanguages') மொழிகளுக்கு இடையே உள்ள பொருத்தங்களையும் வேறுபாடுகளையும் தீர்மானிக்க இயலும்; (ஊ) மொழிபெயர்ப்பு செய்யும் போது புதிய பனுவல்களைச் சேர்த்தும் பனுவல்களிலிருந்து கற்றும் புதுச்சொல்லாக்கங்களைக் (neologisms) கையாளத் தரவுத்தளங்களை எளிதில் மேம்படுத்த இயலும்; (எ) மொழிபெயர்ப்பு நிபுணத்துவம் சிறந்த மொழிபெயர்ப்பாளர்களைப் 'போலச்செய்து' பெறப்பட்டது; அதாவது அகராதியில் பெரும்பாலும் காணப்படாத கலவைத்தன்மையான சூழல்சார் தகவல் வளத்தைப் பயன்படுத்தி பெறப்பட்டது.

எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் கருத்துரு UMIST-ATR திட்டத்தில் காணப்படுகின்றது.

3.20.3. ஐபிஎம்-இல் புள்ளியியல் அடிப்படையில் இயந்திர மொழிபெயர்ப்பு

பெரும்பாலான ஒழுங்குமுறைகளும் நெறிமுறைகளும் வாக்கியங்கள் மற்றும் பனுவல்களின் மொழியியல் பகுப்பாய்வையும் உருவாக்கத்தையும் விளக்கும். இது மரபு

‘மொழியியல் அடிப்படை’யிலான ஒழுங்குமுறைகளுக்கு மட்டுமன்றி (Systran, SUSY, Ariane, Eurotra, METAL) வழக்கத்திற்குச் சிறிது மாறான Météo, Roetta, DLT என்ற ஒழுங்குமுறைகளுக்கும் அறிவு அடிப்படையிலான மற்றும் எடுத்துக்காட்டு அடிப்படையிலான அணுகுமுறைகளுக்கும் உண்மையாகும். 1960களின் தொடக்கத்திலிருந்து இயந்திர மொழிபெயர்ப்பு ஆய்வை ஆதிக்கம் செலுத்திய அனுமானங்களிலிருந்து தீவிர மாற்றம் யார்க் டவுண் ஹைட்ஸில் ஐபிஎம் ஆய்வுப் பரிசோதனைக் கூடங்களின் ஆய்வால் உருப்படுத்தம் செய்யப்படுகின்றது; இந்த மொழிபெயர்ப்பு ஒழுங்குமுறை முற்றிலும் புள்ளியியல் நுட்பங்களின் அடிப்படையில் அமையும்.

புள்ளியியல் பகுப்பாய்வின் பயன்பாடு தொடக்ககால இயந்திர மொழிபெயர்ப்பில் பழக்கமின்றி இருந்தது என்று கூற இயலாது. கணினிசார் செயல்முறைக்கு மொழியைப் பற்றிய சமகால அறிவு போதுமானதல்ல என்ற அனுமானத்தின் மீது புள்ளியியல் பகுப்பாய்வு மொழியியல் தரவின் தானியக்க வகைப்பாட்டிற்கு முதன்மைக் கருவியாகப் பயன்படுத்தப்பட்டது. இந்தப் பயன்பாடு தற்காலம் வரை தொடர்கின்றது; பல ஆய்வு திட்டங்கள் புள்ளியியல் தரவை விதிகள் எழுதுவதற்கும் வழக்கமானவைகளின் உருவாக்கத்திற்கும் வழிகாட்டப் பயன்படுத்துகின்றன. ஐபிஎம் திட்டத்தின் தனித்துவம், பகுப்பாய்விற்கும் உருவக்கத்திற்கும் புள்ளியியல் நுட்பங்களை முழுக் கருவியாகப் பயன்படுத்துவது ஆகும். பேச்சு புரிதலிலும் பகுப்பாய்விலும் புள்ளியியல் அடிப்படையான அணுகுமுறைகளின் அதிகரிக்கும் அதிநவீன பயன்பாடுகளால் இது சாத்தியமாக்கப்பட்டது.

ஐபிஎம் ஆராய்ச்சி, ஆங்கிலத்திலும் பிரஞ்சிலும் பாராளுமன்ற விவாதங்களைப் பதிவுசெய்த கனடியன் ஹான்ஸர்டின் (Canadian Hansard) பெரிய தரவுத்தொகுதிகள் அடிப்படையிலானதாகும். பரிசோதனைக்கான தரவுத்தொகுதி ஒவ்வொரு மொழியிலும் 40,000 இணை வாக்கியங்கள் அடங்கியதாகும். இந்நெறிமுறையின் சாரம், இரு மொழிகளின் வாக்கியங்களின் இணையான வரிசைப்படுத்தலாகும் மற்றும் ஒரு மொழியின் ஒரு வாக்கியத்தில் ஏதாவது ஒரு சொல், மற்ற மொழியில் மொழிபெயர்க்கப்பட்ட வாக்கியங்களில் இரண்டு, ஒன்று அல்லது பூஜிய சொற்களுடன் பொருந்தும் சாத்தியமான கணக்கீடாகும். ஒவ்வொரு ஆங்கில வாக்கியங்களின் பைகிராம்களை (இரண்டு தொடர்ச்சியான சொற்கள்) ’நிகரான’ (equivalent)

பிரஞ்சு வாக்கியங்களின் பைகிராம்களுடன் (bigrams) பொருத்துவதால் சாத்தியங்கள் கணக்கிடப்படுகின்றன.

இரண்டு கண சாத்தியங்கள் கணக்கிடப்பட்டுள்ளன. முதலில் ஒவ்வொரு தனித்தனியான ஆங்கிலச் சொல்லுக்கும் ஒரு கண பிரஞ்சு சொற்களுடன் பொருந்தும் இதன் சாத்தியங்கள்; எடுத்துக்காட்டாக, *the* பிரஞ்சு *le* உடன் .610 சாத்தியத்துடம் *la* உடன் .178 சாத்தியத்துடனும் *l'* உடன் .083 சாதியத்துடனும் *les* உடன் .023 சாத்தியத்துடனும் *ce* உடன் .013 சாத்தியத்துடனும் il உடன் .012 சாத்தியத்துடனும் பொருந்தும் போன்றன. இரண்டாவது இரண்டு, ஒன்று அல்லது பூஜியம் பிரஞ்சு சொற்கள் ஒரு தனி ஆங்கிலச் சொற்களுடன் பொருந்தும் என்ற சாத்தியங்கள்; எடுத்துக்காட்டாக, *the* ஒரு பிரஞ்சு சொல்லுடன் .871 சாத்தியத்துடனும் பூஜியத்துடன் .124 சாத்தியத்துடனும் இரண்டுடன் .004 சாத்தியத்துடனும் பொருந்தும். *not* என்பது இரண்டு பிரஞ்சு சொற்களுடன் .758 சாத்தியத்துடன் பொருந்தும்; இவை பெரும்பாலும் *ne* (.460) மற்றும் *pas* (.469) என்பனவாகும்; *plus* (.002) மற்றும் *jamis* (.002) என்பதுடன் குறைந்த சாத்தியமாகும்; *not* உடன் பிற பொருத்தங்கள் *non* (.024), *pas du tout* (.003), மற்றும் *faux* (.003) போன்றன.

அணுகுமுறைகளின் திறன் பிரஞ்சிலிருந்து ஆங்கிலத்திற்கு மொழிபெயர்ப்பால் மதிப்பீடு செய்யப்பட்டது. சொற்றொகை அதிக நிகழ்வெண் உள்ள ஆங்கிலச் சொற்கள் 1000க்கும் அவற்றின் பொருத்தம் அதிக நிகழ்வெண் உள்ள 1,700 பிரஞ்சு சொற்களுக்கும் எல்லைப்படுத்தப்பட்டது. மொழிபெயர்ப்பு மாதிரி ஹன்ஸர்ட் தரவுத்தொகுதியில் எங்கிருந்தோ 73 பிரஞ்சு வாக்கியங்களின் மீது பரிசோதிக்கப்பட்டது. விளைவுகள் பின்வருமாறு வகைப்படுத்தப்பட்டுள்ளன: (அ) சரியானவை (ஹன்ஸர்ட் மொழிபெயர்ப்புக்குச் சமம்), (ஆ) மாற்று (ஒரே அர்த்தம் ஆனால் சிறுது வேறுபட்டச் சொற்கள்), (இ) வேறுபட்டவை (முறையான மொழிபெயர்ப்பு ஆனால் ஹன்ஸர்ட் மொழிபெயர்ப்பை ஒத்த அர்த்தத்தைத் தரவில்லை), (ஈ) தவறு (புரியக்கூடிய விளைவு ஆனால் பிரெஞ்சின் மொழிபெயர்ப்பல்ல), மற்றும் (உ) இலக்கணத்தவறானது (அர்த்தம் தெரிவிக்கப்படவில்லை). சில எடுத்துக்காட்டுகள் அட்டவணை 18.1. இல் காட்டப்பட்டுள்ளன.

exact		Ces amendements sont certainement necéssaries
	Hansard	These ammendments are certainly necessary
	IBM	These ammendments are certainly necessary

alternative		C’est pourtant très simple
	Hansard	Yet it is very simple
	IBM	It is still simple
different		J’ai reçu cette demande en effet
	Hansard	Such request was made
	IBM	I have received this request in effect
wrong		Permettez que je donne un exemple à la Chambre
	Hansard	Let me give the House one example
	IBM	Let me give an example in the House
ungrammatical		Vous avez besoin de toute l’aide disponible
	Hansard	You need all the help you can get
	IBM	You need of the whole benefits available

5% ‘சரியானது’ வகையில் இருந்தாலும், மொழிபெயர்ப்புகள் முதல் மூன்று வகைகளுக்குள் (சரியானவை, மாற்று, வேறுபட்டவை) வந்தால் ‘நியாயமானவை’ என்று கருத்தப்பட்டன. இந்த அளவுகோலில் ஒழுங்குமுறை 48% வெற்றியுடன் செயல்பட்டது. பெரிய தரவுத்தொகுதியுடன் (10% ஹன்ஸர்ட் மட்டும் தான் பயன்படுத்தப்பட்டது) முன்னேற்றங்கள் எதிர்பார்க்கப்பட்டன. வாக்கியங்களைத் தொடர்களாகச் சாத்தியம் அடிப்படையில் கூறிட்டும், டிரைகிராமையும் பைகிராமையும் பயன்படுத்தியும், ஒன்றாகக் குழும திரிபுசார் உருபனியல் தரவை உட்பத்தியும் எடுத்துக்காட்டாக *tall*, *taller*, *talest* மற்றும் *va*, *vais*, *vont*.

முன்மொழியப்பட்ட கூறிடுதல் மற்றும் உருபனியல் தரவை உட்படுத்தல் புள்ளியியல்சார் அணுகுமுறைகளுக்கு உள்ளார்ந்த எல்லைகள் உண்டு என்பதைத் தெரிவிக்கின்றது. இருப்பினும் இந்த ஆய்வின் முக்கியத்துவம் மொழியியல் பகுப்பாய்வுக்கு மூலவளம் இல்லாமல் இருமொழிய ஒருதிசை இயந்திரமொழிபெயர்ப்புக்குச் செல்வது எந்தளவுக்குச் சாத்தியமானது என்பதை நிரூபிக்கின்றது. இருப்பினும் விளைவுகள் ஒரு குறிப்பிட்ட தரவுத்தொகுதிக்குப் பக்கச்சார்பானது என்பதை நினைவு கூறவேண்டும். எடுத்துக்காட்டாக, இந்த பனுவல்களில் hear என்பதன் மொழிபெயர்ப்பு எப்போதும் bravo (சாத்தியம் .992)ஆகும்; அதாவது கன்னடா பாராளுமன்றத்தில்

Hear, hear! என்பது *Bravo!* என்பதுடன் பொருந்தும்; ஆனால் சாதாரண மொழிபெயர்ப்பு *entendre* குறைந்த சாத்தியம் .005ஐக் கொண்டுள்ளது.

புள்ளியியல் அடிப்படையிலான நுட்பங்கள் பல எதிர்கால இயந்திர மொழிபெயர்ப்புத் திட்டங்களின் பண்பாக அமையும்; ஆனால் ஐபிஎமின் முற்றிலும் புள்ளியியல் அடிப்படை என்ற நிலையைப் பின்பற்றுவார்களா என்பது நிச்சயமற்றது. தற்போது மொழியியல் தரவு மற்றும் நெறிமுறைகள் எந்த நடைமுறை இயந்திர மொழிபெயப்பு ஒழுங்குமுறையின் மையத்திலிருக்கும் என்ற அனுமானம் நிலைத்துநிற்கின்றது. (1992இல் ஹட்சின்ஸ் மற்றும் சோமெரால் இப்படிக் குறிப்பிடப்பட்டுள்ளது. தற்போது அதாவது 2000ங்களில் மொழியியல் தரவுச்சார்பிலிருந்து விலகி முற்றிலும் புள்ளியியல் அடிப்படையிலான ஒழுங்குமுறைகள் முயற்சிக்கப்பட்டு வருகின்றன.)

3.20.4. துணைநிலைமொழி மொழிபெயர்ப்பு : டைடஸ்

ஒரு குறிப்பிட்ட பனுவல் தரவுத்தொகுதிக்கு இயந்திர மொழிபெயர்ப்பைக் கட்டுப்படுத்துவது துணைநிலைமொழி அணுகுமுறையின் தீவிர மாறுபாடு என்று கருத இயலும். குறிப்பாகத் தரவுத்தொகுதி பெரிதாக இருந்தால்தான் இது நியாயமானதாகும். இது விமானப்போக்குவரத்துக் கையேடுகளுக்கான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதற்கு வேண்டப்பட்ட டாம் குழுவுக்குப் பொருத்தமானதாக இருந்தது. இருப்பினும் துணைநிலைமொழி ஒழுங்குமுறை Météoவின் வெற்றியைத் திருப்பச்செய்யப் இயலவில்லை.

நடைமுறையில், பரிசோதனையாக இருந்தாலும் சரி வணிநோக்காக இருந்தாலும் சரி பெரும்பாலும் எல்லா இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் ஒரு குறிப்பிட்ட பொருண்மைக் களத்திற்கு எல்லைப்படுத்தப்படிருந்தன. இதை சிஸ்ட்ரானின் பல செயல்படுத்தங்களில் காணலாம்: பான் அமெரிக்க உடல்நல நிறுவன ஒழுங்குமுறைகள் (மருத்துவம் மற்றும் பொதுமக்கள் உடல்நல ஆவணங்கள்), யுரெட்ரா மூலமுன்வகை (தகவல் நுட்பம்), மெடல் (தொழில் நுட்ப ஆவணங்கள்), சிம்யு ஒழுங்குமுறை (தனிநபர் கணினி கையேடுகள்), கணினி தொழிநுட்பம், மின்னணுப் பொறியியல் இவற்றின் மொழிபெயர்ப்புகளில் கவனக்குவிப்பு செய்திருந்த ஜப்பான் கணினி நிறுவனங்களிலிருந்து பல நுண்கணினி அடிப்படையிலான ஒழுங்குமுறைகள். சில கூடுதலான எடுத்துக்காட்டுகள் பின்னர் வரும். இருப்பினும் இந்த ஒழுங்குமுறைகள் 'துணைநிலை மொழி ஒழுங்குமுறைகள்' என

வகைப்படுத்தப்படவில்லை; ஏனென்றால் அவைகள் ஒரு குறிப்பிட்ட பாடப்பொருள்களுக்குக் கட்டுப்படுத்த வடிவமைக்கப்படவோ நோக்கமாகவோ கொண்டவை அல்ல. அவற்றின் தற்போதைய வரையறை எல்லைகள் தற்காலிகமாகக் கருதப்படுகின்றன; பிற பாடப் புலங்களுக்கு நீட்சிசெய்வது எதிர்பார்க்கப்படுகின்றது. இதற்கு முரணாக, துணைநிலை மொழிபெயர்ப்பு ஒழுங்குமுறைகள் ஒப்புருச் சொன்மை, மொழிபெயர்ப்புசார் பொருண்மை மயக்கம் மற்றும் அமைப்புப் பல்வேறானவை சிறப்பாக ஒரு குறிப்பிட்ட பாடப்பொருளுக்காக அல்லது பனுவல் வகைக்காக உருவாக்கப்பட்டவை. முன்மாதிரி அமைப்பு Météo ஆகும்; மற்றொரு எடுத்துக்காட்டு டைடஸ் (TITUS).

டைடஸ் ஒழுங்குமுறை, நெசவுத் தொழிலுக்காக நிகழ்நிலைத் தரவுதளத்தில் உள்ள சுருக்க உரைகளின் பன்மொழிய நடத்துமுறைக்காக பிரஞ்சு நெசவு நிறுவனத்தால் (Institut Textile de France) வடிவமைக்கப்பட்டதாகும். 1970இல் முதலில் நிறுவப்பட்ட ஒழுங்குமுறை அதன் நான்காவது வகைபேதத்தில் (1992இல்) இருக்கின்றது. சுருக்க உரைகள் பின்வரும் நான்குமொழிகளில் பனுவல்களை உருவக்க இயலும் உருப்படுத்தங்களில் சேகரிக்கப்பட்டுள்ளன: ஆங்கிலம், பிரஞ்சு, . ஜெர்மன் மற்றும் ஸ்பானிஷ். சுருக்க உரையை இந்த நான்கில் ஏதாவது மொழியில் உள்ளீடு செய்ய இயலும்; அவை கட்டுப்படுத்தப்பட்ட தொடரியலாலும் கட்டுப்படுத்தப்பட்ட சொற்றொகையாலும் நெசவுத்தொழில் துணைநிலை மொழியின் நிலைபேறுபெற்ற கலைச்சொற்களாலும் வடிவமைக்கப்பட்டுள்ளன. கட்டுப்படுத்தப்பட்ட தொடரியல் அடிப்படையான தொடர் வகைகளின் நிரலை தீர்மானிக்கின்றது: எழுவாய் பெயர்த்தொடர், சூழ்நிலைப் பெயர்த்தொடர், வினைத்தொடர், நிரப்பிப் பெயர்த்தொடர், முன்னுருபுசார் பெயர்த்தொடர் (இங்கு இந்தப் பெயர்த்தொடரில் சில சமநிலைத் தொடர்கள் மற்றும் வினைத்தொடரும் நிரப்பிப் பெயர்த்தொடரும் விருப்பாகும்). தொடரியல், உறுப்புக்களின் நிரல் அடிப்படையிலும் விருப்பு அடிப்படையிலும் பெயர்த்தொடர்கள் மற்றும் வினைத்தொடர்களின் அமைப்பை விளக்குகின்றது. முன் திருத்தக் கட்டத்தில், சாத்தியமான பொருண்மைமயக்கச் சொற்கள் வேறுபடுத்தப்படுகின்றன; எடுத்துக்காட்டாக பிரஞ்சு வினைவடிவம் a என்பது பின்னுருபு a (கடுமையான ஒலி அழுத்தம் தரப்படாத) என்பதிலிருந்து பின் தொடரும் சாய்வுக்கோடால் வேறுபடுத்தப்படுகின்றது. மேலும், அமைபுப் பொருண்மை மயக்கம் (எ.கா. பின்னுருபுத்தொடர்களின் முன்வருகிளவி) நிறுத்தற்குறிகளின் செருகலால்

நீக்கப்படுகின்றன. சுருக்க உரைகள் ஊடாட்டத்தால் உள்ளீடு செய்யப்படுகின்றன; ஒழுங்குமுறை கட்டுப்படுத்தப்பட்ட தொடரியலுடன் இணங்காத்த தொடர்கள் மற்றும் வாக்கியங்களின் மறுவடிவாக்கங்களை வேண்டும் மற்றும் ஒப்புருச்சொற்கள் மற்றும் அமைப்புப் பொருண்மை மயக்கங்கம் இவற்றின் விளக்கவுரையைக் கேட்கும். கட்டுப்பாடுகள் இருந்தபோதிலும் அனுமதிக்கப்பட்ட அமைப்புகள் கவர்ச்சியுள்ளவை. (10) இதை எடுத்துக்காட்டும்.

(10) L'analyse du fluage des fibres de polyéthyléne apès irridiation sous vide monte ; qu'un processus de pontage servient dans la région amorphe ; tandis qu'une coupure de la chaine moléculaire de la région crystalline est observée.

உள்ளீட்டிற்குப் பின் ஒழுங்குமுறை, புகுப்பாய்வு சரியானதா என்பதை அறிய உள்ளீட்டு மொழியின் ஒரு வடிவபேதத்தை (version) உருவாக்கும்; பின்னர் பிற மொழிகளில் வடிவபேதங்களை (versions) உருவாக்கும். எடுத்துக்காட்டாக ஆங்கில உள்ளீட்டு (10)க்குச் சமமான வடிவம் (11) ஆகும்.

(11) The polyethylene fibre creep analysis after irradiation under vacuum shows that a cross-linking process occurs in the amorphous region whereas a molecular chain scission of crystalline region is observed.

உள்ளீட்டின் கட்டுப்பாடு, விரிவான முன் திருத்தம், பகுப்பாய்வின் போதும் பகுப்பாய்வுக்குப் பின்னரும் ஊடாட்டப் பின்னூட்டம் மற்றும் முறைப்படுத்தப்பட்ட துணைநிலை மொழிக்கு வரையறை இவற்றை ஒன்றிணைப்பதால் மொழிபெயர்ப்புகளின் தரம் உறுதி செய்யப்படும். இந்த விதமான கட்டுப்பாடுகளைக் கருத்தில் கொள்ளும் போது டைடஸ் ஒரு உண்மையான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையாகக் கருதப்படவேண்டுமா என்று ஐயுறலாம். டைடஸும் Météoவும் நன்கு வரையறுக்கப்பட்ட சூழல்களில் என்ன சாதிக்க இயலும் என்பதை எடுத்துக்காட்டுகின்றன.

அதிக அளவில் துணைநிலைமொழி ஒழுங்குமுறைகள் உருவாக்கப்படவில்லை என்பது ஆச்சரியத்திற்குரியதாகும். இருப்பினும் ஒப்பீட்டளவில் தற்கட்டுப்படுத்தப்பட்ட சொல்சார் மற்றும் தொடரியல்சார் பொருண்மைக் களத்திற்கு மொழிபெயர்ப்பைக் கட்டுப்படுத்த இயலும் மிகக் குறைவான சூழல்களே இருக்க இயலும் எனத் தோன்றுகின்றது. TAUM குழுவிலிருந்து வந்த Météo ஆய்வாளர்கள் துணைநிலை நெறிமுறைகளைப் பயன்படுத்த ஏற்றதான புலத்தைத்

தேடிக்கொண்டிருந்தனர்; தற்போது சிலர் கால்நடைச் சந்தை அறிக்கைகளில் செயல்படுகின்றனர். ஆய்வுக்கு எடுத்துக்கொண்ட பிற துணைநிலைமொழிகள் வணிகக் கடிதத்தொடர்பு, உணவகம் மற்றும் கருத்தரங்கு முன்பதிவுகள் மற்றும் காவலர் கருத்துப்பரிமாற்றம் இவற்றை உட்படுத்தும்.

3.20.5.ஒருமொழியப் பயன்பாட்டாளர்களுக்கு இயந்திர மொழிபெயர்ப்பு

முன்னர் கருத்துரையாடியது போல் பெரும்பாலான இயந்திர மொழிபெயர்ப்பு ஒழுக்குமுறைகள் (வெளிப்படையாகவோ உட்படையாகவோ) மூலமொழி, இலக்கு மொழி என்ற இரண்டும் அறிந்த நடப்பு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் குறைபாடுகளைச் சரிகட்ட இயலும் பயன்பாட்டாளர்களுக்காக உருவாக்கப்பட்டவை. இருப்பினும் மூலமொழியில் அல்லது இலக்கு மொழியில் பழக்கமில்லாத ஒருமொழியப் பயன்பாட்டாளர்கள் இயந்திர மொழிபெயர்ப்பைப் பயன்படுத்துவது சாத்தியமானதாகும். சிஸ்ட்ரான் போன்ற தொகுதி ஒழுங்குமுறைகளிருந்தான வெளியீடு பொருண்மைப் புலத்தில் சிறப்பு நிபுணர்களுக்கு மதிப்புளதாகும்; ஏனென்றால் இந்த நிபுணர்கள் மூலமொழியை அறிந்திருக்கத் தேவை இல்லை; அவர்கள் மொழிபெயர்ப்புச் செயல்பாட்டில் உதவியாக இருப்பதை எதிர்பார்க்க இயலாது. ஆனால் நாம் செயல்பாட்டின் மறுமுனையில் உள்ள பயன்பாட்டாளர்களைக் கருத்தில் கொண்டால் அதாவது பனுவலின் உள்ளீட்டைக் கருத்தில் கொண்டல் சுவாரஸ்யமான சாத்தியங்கள் உள்ளன. இவற்றில் சில ஊக்கமான புலனாய்வின் கீழ் உள்ளன; குறிப்பாக, பயனர் அறியாத மொழியில் மொழிபெயர்க்கப்பட உள்ள பனுவல்களின் ஊடாட்ட கூட்டமைவுக்கான (interactive composition) ஒழுங்குமுறைகள்.

உமிஸ்டில் (University of Manchester Institute of Science and Technology (UMIST)) ஒருமொழிய பயனர்களுக்கு வேண்டி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் முன்னோடியான செயல்பாடு என்டிரானின் (Ntran) உருவாக்கத்தில் மேற்கொள்ளப்பட்டது. இது ஆங்கில மூலப் பனுவலின் ஊடாட்ட பொருண்மை மயக்கநீக்கத்தை உட்படுத்திய ஆங்கிலம்-ஜப்பானியமொழி ஒழுங்குமுறை ஆகும்; இதில் கூடுதலான பொருண்மை மயக்கங்கள் சொல் மாற்றத்தின் போது எழும்; இது ஜப்பானிய மொழி அறியாத பயனாளி பொருண்மை மயக்கங்களைத் தீர்ப்பதை எதிர்பார்க்கும் ஊடாட்டத்தின் கட்டம் ஆகும் (அதாவது இலக்குமொழி வேறுபாடுகளின் ஆங்கில பொழிப்புரை அடிப்படையில்). ஜப்பானிய மொழி உருவக்கம் முற்றிலும் தானியக்கமாகும். இந்த மூலமுன்மாதிரியின் ஒத்தறி வெற்றியால் ஊக்குவிக்கப்பட்ட

உமிஸ்ட் ஆய்வாளர்கள் ஒருமொழிய பயனருக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறையின் உருவாக்கத்தை பொதுக் கருத்தாகக் கொண்ட பல பிற ஆய்வுத்திட்டங்களில் ஈடுபட்டனர்.

பிரிட்டிஷ் டெலிகாமால் (British Telcom) நிதி நல்கப்பட்ட இம்மாதிரியான ஒரு ஆய்வுத்திட்டம் இயந்திர மொழிபெயர்ப்புத் ஒழுங்குமுறையை உருவாக்குவதை நோக்கமாகக் கொண்டது; இது விருப்பத்தேர்வுகளின் பட்டியல் வழி வழிகாட்டி பழக்கமில்லாத வெளிநாட்டு மொழியில் வணிகக் கடிதங்களை (எ.கா. புகார், சலுகை, விசாரணை போன்றன) எழுத பயனர்களுக்கு உதவும். இந்த ஒழுங்குமுறை வணிகக் கடிதங்களின் குறிப்பிட்ட வகைகளோடு பொருந்தும் 'முறையான படிவ' (proforma) பனுவல்களின் உத்தி அடிப்படியிலாகும். முறையான படிவப் பனுவல்கள் கணத் தொடர்கள், பெயர்கள், நாட்கள், முகவரிகள் போன்றவற்றின் பொருந்துமிடங்களைக் கொண்ட வார்ப்புருக்கள் (templates) ஆகும்; அவை பயனர்களால் அவர்களின் சொந்த மொழியில் உள்ளீடுசெய்யப்படவேண்டும். முறையான படிவம் முழுமையுற்றால், 'முன் மொழிபெயர்க்கப்பட்ட' பனுவல் துண்டுகளின் தரவுத்தளத்துடன் பயனர்களால் தரப்பட்ட தகவலை ஒப்பிட்டு பனுவலின் பன்மொழிய நிகரன்களை ஒழுங்குமுறை உருவாக்க இயலும். இந்த ஒழுங்குமுறையின் கவர்ச்சி, இலக்கு மொழியின் வார்ப்புருக்கள் மனித மொழிபெயர்ப்பாளர்களால் எழுத்தப்பட்ட நடை அடிப்படையில் பொருத்தமான பனுவல்கள் அடிப்படையிலானதாகும் என்பதால் பனுவல் தரப்பட்ட பொருண்மைக் களத்திற்குள் கண்டிப்பாக இருப்பதுவரை மிகத்தரமான வெளியீடை உத்தரவாதம் செய்ய இயலும். இதுபோன்ற அணுகுமுறை மலாய் மொழியில் அலுவலகக் கடிதங்களை வரைவு செய்ய உதவும் மலேசியாவில் உருவாக்கப்பட்ட ஒழுங்குமுறையிலும் காணப்படுகின்றது.

மற்றொரு அணுகுமுறையும் உமிஸ்டில் உருவாக்கப்பட்டுள்ளது; இது 'மூலப் பனுவல்' (source text) என்பதன் இடத்தில் ஒழுங்குமுறை-பயனர் (system-user) ஊடாட்ட உத்தியை கூடுதல் எடுத்துச் செல்கின்றது. ஜப்பானிய எடிஆர் (Advanced Telecommunications Research (ATR)) ஆய்வுக்கூடங்ளுடன் இணைந்து மேற்கொள்ளப்படும் இந்த ஆய்வுத்திட்டத்தில் ஒரு உரையாடல் மொழிபெயர்ப்பு ஒழுங்குமுறை (dialogue translation system) உருவாக்கப்பட்டுள்ளது. பொருண்மைக்களம், ஜப்பானிலுள்ள கருத்தரங்க அலுவலகத்திற்கும் ஆங்கிலம் பேசும் கேள்விகேட்பவருக்கும் இடையில் நிகழ்நிலை உரையாடல் (online conversation) (இறுதியாக தொலைபேசியில், ஆனால் விசைப்பலகை உரையாடல் தான் ஊடகம்)

ஆகும். இந்த ஒழுங்குமுறை, இரண்டு உரையாடும் பங்காளிகளுக்கு இடையில் ஆங்கிலத்திற்கும் ஜப்பான் மொழிக்கும் இடையில் அவர்தம் உரையாடலை மொழிபெயர்க்கும் இடைத்தரகராக செயல்படும் என்பதுதான் உத்தியாகும். உரையாடல் மொழிபெயர்ப்பு குறிப்பாகக் கடினமான செயல்படாகும்; ஏனென்றால் சொற்கள் விடுபட்ட வாக்கியங்களின் அதிக நிகழ்வெண்ணும் அரைகுறை கூற்றுகளும் தவறாக உருவாக்கப்பட்ட கூற்றுகளும் முற்சுட்டு (anaphora) நேரடிச்சுட்டுக் குறிப்புரை (deictic reference) இவற்றின் பயன்பாடு என்பன உரையாடல் மொழிபெயர்ப்பைக் கடினப்படுத்தும். மற்றும் ஒப்புட்டுப்பார்க்கும் போது இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளால் வழக்கமாக மொழிபெயர்க்கப்படும் பனுவல்களின் வகைகளைக் காட்டிலும் உரையாடல் மொழியின் பெரிய பகுதியைக் கொண்டிருக்கும்; இங்கு நேரடி மொழிபெயர்ப்புப் பொருத்தமாக இராது: ஒற்றுமையுள்ள கூற்றுகள் முற்றிலும் வேறுபட்ட உரையாடல் செயல்பாடுகள் உடையவையாய் இருக்கும்; எனவே மொழிபெயர்ப்பு ஒரு கணத்திலிருந்து மற்றொருகணத்தில் வேறுபடும். எடுத்துக்காட்டாக, உரையாடலில் OK என்பது பின்வரும் ஏதாவது தொடர்களைக் குறிப்பிடலாம்.

(12a) I agree with you.

(12b) I can still hear you.

(12c) Let's change the subject now.

(12d) That is good.

உமிஸ்ட் -பிரிட்டிஷ் டெலிகாம் ரிசர்சைப் போல் ஒழுங்குமுறை ஒரு இருமொழிய 'உரையாடல் மாதிரி' இருக்கும்; இம்மாதிரியில் சாத்தியமான கூற்றுக்களைப் பயனர்களின் உள்ளீட்டுக்குப் பொருத்த வேண்டிப் பயனர்களோடு ஊடாடும். இரண்டு சாத்தியமான சூழல் காட்சிகள் உள்ளன. ஒன்றில் முன்மொழியும் கூற்றுக்களைச் தட்டச்சு செய்து பயனர் முனைப்பு எடுப்பர். மற்றொன்றில் உரையாடல் மாதிரி அடிப்படையில் அடுத்த கூற்று எதாக இருக்கவேண்டும் என்று முன்மொழியம் செய்து ஒழுங்குமுறை முனைப்பு எடுக்கும்.

ஒழுங்குமுறைகள் ஒரு தரவுத்தளத்தில் முன் மொழிபெயர்ப்பு செய்த பனுவல் துண்டுகளுடன் உள்ளீட்டுத் தொடர்களைப் பொருத்துவதற்கு இந்த ஒழுங்குமுறைகள் முயற்சி செய்வது வரை, அவை எடுத்துகாட்டு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் இனத்தைச் சாரும். லிடியா (Large Internationalisation of Documents through Interaction with Authors

(LIDIA)) என்ற இலட்சியத் திட்டம் கிரெனோபிள் பல்கலைக்கழகத்தால் முன்மொழியப்பட்டது. அதன் மொழியியல் நுட்பங்களுக்கு இது கேதாவின் (GETA's) நன்றாகப் பரிசோதிக்கப்பட்ட ஒழுங்குமுறை அரினே ஒழுங்குமுறை (Ariane system) அடிப்படையில் ஆனதாகும். இதன் நோக்கம் ஒரு ஊடாட்ட ஒழுங்குமுறை ஆய்வாளர்களை அவர்களின் சொந்தப் பனுவல்களை அவர்களின் சொந்த மொழியிலிருந்து (பிரஞ்சிலிருந்து) மற்றொரு அறியாத மொழியில் (ஜெர்மன் அல்லது ரஷ்யன்) மொழிபெயர்க்கவும் மற்றும் இலக்கு மொழிப் பனுவலின் மொழிபெயர்ப்பை மூலமொழிக்குத் திரும்ப மொழிபெயர்த்துச் சரிபார்க்கவும் இயலச்செய்கின்றது. உமிஸ்ட் திட்டங்களைப் போல் அல்லாமால் இது முன் மொழிபெயர்க்கப்பட்ட பனுவல் துண்டுகள் அடிப்படையிலானது அல்ல; ஆனால் அறிமுகமான இயந்திர மொழிபெயர்ப்பு முறையில் பெரிய எழுதப்பட்ட பனுவல்களை மொழிபெயர்க்கும்.

3.20.6.பேச்சு மொழிபெயர்ப்பு: பிரிடிஷ் டெலிகாம் மற்றும் எடிஆர்

அண்மைக்காலங்களில் பேச்சுத் தொழில்நுட்பத்தின் முன்னேற்றங்கள் மொழிபெயர்ப்பு ஒழுங்கு முறைகளைப் பேச்சு அறிதலுடனும் பேசு உருவக்கத்துடனும் ஒன்றிணைப்பதை ஆய பல எண்ணிக்கையிலான ஆய்வாளர்களை ஊக்கபடுத்தியுள்ளன. நாம் புள்ளியியல் அடிப்படையிலான மொழிபெயர்ப்பு ஒழுங்குமுறையில் பேச்சு ஆய்விலிருந்து நெறிமுறைகளின் பயன்பாட்டை முன்னர் பார்த்தோம். நாம் இரண்டு திட்டங்களைச் சுருக்கமாக விளக்குகின்றோம்; இதில் உள்ளீடும் வெளியீடும் பேச்சாகும்; ஆனால் மொழிபெயர்ப்பு மரபு மொழியியல் அடிப்படை அணுகுமுறைகள் ஆகும்.

பிரிட்டிஷ் தொலைத்தொடர்பு ஆய்வு ஆய்வுக்கூடங்களில் (British Telecom Research Laboratories) பரிசோதனை, தொலைபேசி வணிகக் கருத்துப்பரிமாற்றம்/தொடர்பு (Telephonic Business Communication) என்ற மிகக் கட்டுப்படுத்தபட்ட பொருண்மைக் களத்தில் நிலைபெறு தொடர்களுக்கு இணையாகப் பேச்சுச் சொற்களைப் பொருத்துவதன் அடிப்படையில் அமைந்தது. கட்டுப்பாடு தற்போதைய பேச்சுப் புரிவானின் கடுமையான செயல்பாட்டு எல்லைகளால் வேண்டப்பட்டது. எடுத்துக்காட்டாக (13)இல் தரப்பட்டுள்ள மூன்று தொடர்களில், மூன்று முக்கியச் சொற்களான you, speek, I என்பன தனித்தன்மையாம அடையாளம் காணலுக்குப் போத்மானதாகும்.

(13a) Who do you want to speak of?

(13b) I cannot hear you.

(13c) My I speak to Mr. Smith please>

400 வாக்கியங்கள் கொண்ட தொடர் நூலிலிருந்து மிக பயனுள்ள ஆங்கில முக்கியச் சொற்கள் the, a, I , you, to, room, is hotel, for, of என்பனவாகும். ஒற்றுமையுள்ள கணங்கள் பிரஞ்சு, ஜெர்மன், ஸ்பானிஷ் ஆகிய மொழிகளில் காணப்படுகின்றன. வெளிப்படையாக, நேரங்கள், விலைகள் போன்ற மாறி தனிமங்களை (variable elements) ஒரே வழியில் கையாள இயலாது; ஒழுங்குமுறை மூன்றுகட்டங்களை உட்படுத்தும்: ஒரு தொலைபேசி அழைப்பாளர் பேச்சு அறினானில் ஒரு தொடரை, அதன் ஒவ்வொரு சொல்லையும் தெளிவாக விளக்கி மற்றும் ஒவ்வொருக்கு சொல்லுக்கு இடையிலும் இடைநிறுத்தம் செய்துப் பேச்சு அறிவானில் உள்ளீடுசெய்வார்; கணினி ஒரு தொடரைத் தெரிந்தெடுத்து ஏதாவது மாறியைச் (variable) செயலாக்கம் (process) செய்யும்; விருப்பதேர்வு செய்யப்பட்ட தொடர் திரையில் காட்சிப்படுத்தப்படும். ஒரு தொடராக ஏற்றுக்கொள்ளப்பட்டால் சேகரிக்கப்பட்ட மொழிபெயர்ப்பு கடத்தப்படும்; இறுதிச் செய்தி ஏற்பவர் (கேட்பவர்) முனையில் செயற்கைப் பேச்சின் வெளியீடாகும். பிரிட்டிஷ் தொலைத்தொடர்பு ஆய்வாளர்கள் தற்போது (1992இல்) பிரஞ்சு மற்றும் பிரிட்டிஷ் காவலர் படைகளால் பயன்படுத்துவதற்குக் காவலர் செய்திகளின் ('Policespeak') துணைநிலை மொழியின் மீது ஆய்வை உட்படுத்திய இருமொழிய பேச்சுக் கருத்துப்பரிமாற்ற ஒழுங்குமுறையை ஆய்கின்றனர்.

தொலைபேசி மொழிபெயர்ப்பில் ஒரு இலட்சியத்திட்டம் எடிஆர் பரிசோதனைக் கூட்டத்தில் நடைபெறுகிறது (1992இல்). இந்த நீண்ட காலத் திட்டம் ஆங்கிலத்திற்கும் ஜப்பானிய மொழிக்கும் இடையில் பேச்சு உரையாடல் மொழிபெயர்ப்பைச் செய்வதற்குரிய ஒழுங்குமுறை ஆகும். உரையாடலின் பேச்சு அறிதல், பேச்சு உருவாக்கம் மற்றும் தானியக்க மொழிமாற்றம் இவற்றை உள்ளடக்கும். தற்போதைய பேச்சு அறிவான்கள் பொதுவாகச் சொற்களுக்கு இடையில் இடைநிறுத்தலை வேண்டும்; மேலும் அவை குறிப்பிட்ட தனிநபர்களுக்காக மாற்றி அமைக்கப்பட்டுள்ளது. தொடர்ச்சியான உரையாடலின் பேசுபவர் சுதந்திரமான பேச்சு அறிவான் பிச்சு, அழுத்தம் மற்றும் காலநீட்சி போன்ற மீக்கூறு தகவல்களை உட்படுத்தும் அசை எல்லைகளின் தெரிகை, ஒலியன்கள் மற்றும் சொற்களின் எல்லைகளை அடையாளம் காண பகுப்பானுடன் ஒருங்கிணப்பு, சாத்தியமான பொருள்கோள்களைச் குறுக்கும்

பொருண்மையியல் மற்றும் பயன்வழியியல் தகவல்களின் பயன்பாடு போன்றவற்றை உட்படுத்தும் கருதப்பட்ட சவால்களை முன்வைக்கின்றது. தானியக்கப் பேச்சு உருவாக்கத்தின் தேவைகளும் அதே அளவுக்கு சவால் உள்ளதாகும். தற்போதைய பேச்சு உருவாக்கிகளை ஒப்பிடுகையில் மீக்கூறு, திரிபு இவற்றைக் கணக்கில் எடுக்கும் மற்றும் அதிக அளவில் தனிப்பட்ட தன்மையுடைய உயர்ந்த தரமுள்ள, அதிக இயற்கை தேவைப்படுகின்றது.

உரையாடலின் (dialogue) மொழிபெயர்ப்பைப் பொறுத்தவரையில் இது இயந்திர மொழிபெயர்ப்பின் புதிய களமாகும்: பேச்சு உரையாடல் (speech dialogue) எழுத்துப் பனுவலிலிருந்து (written text) சொற்றொகை மற்றும் இலக்கணம் இவற்றில் வேறுபடும்; ஆனால் முழுமையுறாத கூற்றுகள் (incomplete utterances), தவறான தொடக்கங்கள் (false starts), விடுபாடுகள் (ellipses), கூறப்படாத உட்குறிப்பீடுகள் (unstated implications) போன்றவை பொதுவானதாகும். எடிஆர் ஆய்வுத்திட்டம் பேச்சுச் செயல்கள் (speech acts), உரையாடல் மாற்றம் (dialogue shifting), ஜப்பானிய மொழி மரியாதை வழக்குகள் (Japanese honorifics), அனுமானிக்கும் செயற்பாட்டுச் சொற்கள் (inferring function words), விடுபட்ட எழுவாய்கள் (omitted subjects), (எடுத்துக்காட்டு-அடிப்படையிலான நெறிமுறைகளைப் பயன்படுத்துவதை உள்ளடக்கிய) மரபுத்தொடர் வெளியீட்டை உருவாக்குதல் (producing idiomatic output), முன்னர் கூறியபடி தற்போது ஆய்வு உலகளாவிய கருத்தரங்கு அலுவலகத்தின் கருத்துப்பரிமாற்றச் சூழலில் கவனக்குவிப்புச் செய்யும். ஒரு செயல்பாட்டு மூலமுன்வகை வருகிறதோ இல்லையோ, எடிஆரில் அடிப்படை ஆய்வு பொதுவாக கணிசமான அளவில் இயந்திர மொழிபெயர்ப்புக்குப் பங்களிப்பு செய்துள்ளது; எதிர்காலத்திலும் பங்களிப்பு செய்யும்.

3.20.7. தலைகீழ் இலக்கணம் (Reversible grammar)

இயந்திர மொழிபெயர்ப்பின் மொழியியல் அடித்தளங்கள் பல திசைகளில் தொடர்ந்து ஆயப்பட வேண்டியுள்ளது. சொல்செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG), பொதுமையாக்கப்பட்ட தொடரியல் இலக்கணம் (Generalized Phrase Structure Grammar (GPSG), ஆளுகை கட்டுறவு கோட்பாடு (Government and Binding), முறையான செயல்பாட்டு இலக்கணம் (Systemic-Functional Grammar), வகைப்பாட்டு இலக்கணம் (Categorical Grammar), சூழல் பொருண்மையியல் (Situation Semantics), தர்க்க இலக்கணங்கள் (Logical grammars) போன்ற கோட்பாட்டு மொழியியலின் அண்மைக்கால முன்னேற்றங்களின் பயன்பாட்டிலும்

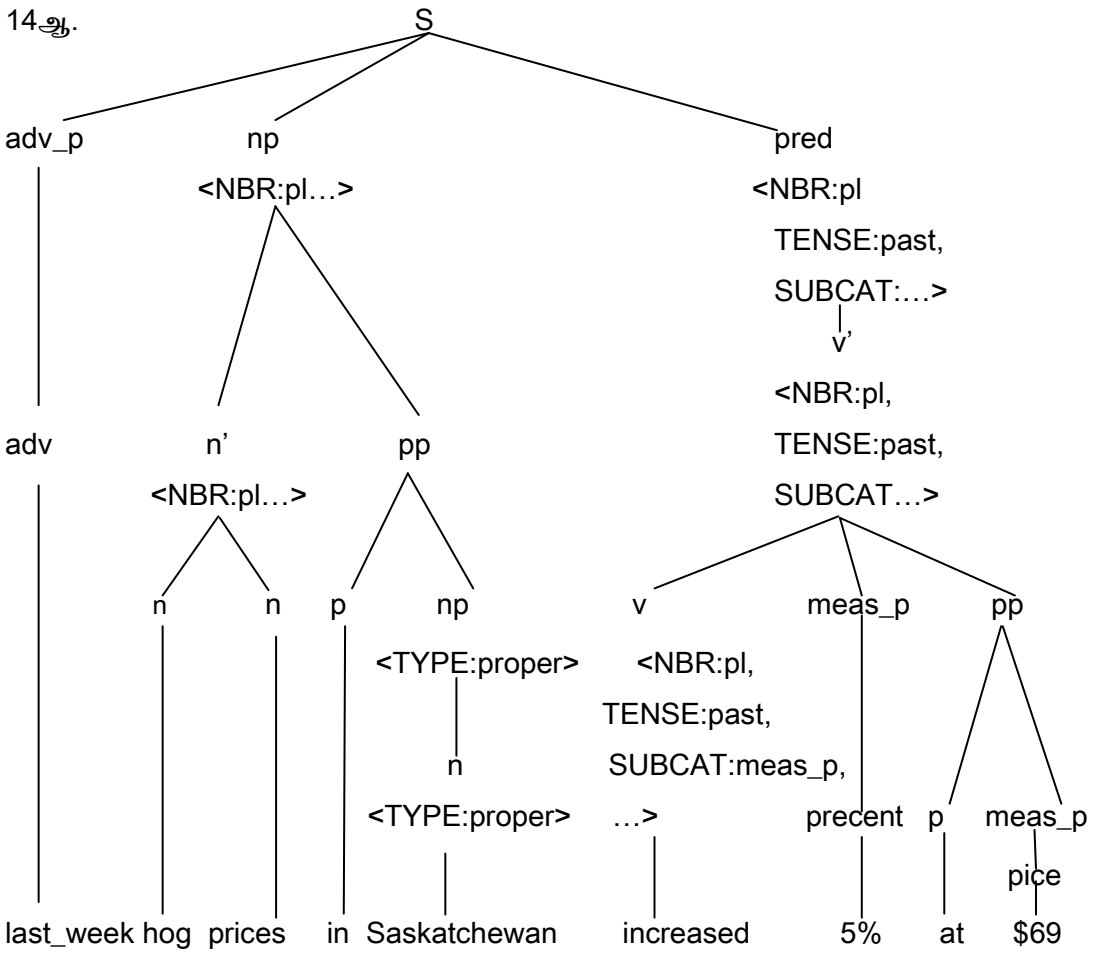
செயல்படுத்தத்திலும் அதிக கவனம் செலுத்தப்படுகின்றது. தற்போது பிற முக்கியமான தலைப்புகள் கூட்டமைவையும் (compositionality) மீளுமையையும் (reversibility) உட்படுத்தும். குறிப்பாக இலக்கணங்களின் மீளுமை பல ஆய்வுத்திட்டங்களின் நோக்கமாய் இருந்தது; ஒரு மாதிரி எடுத்துக்காட்டு கனடாவில் க்ரிட்டர் (CRITTER) ஆய்வுத்திட்டம் ஆகும்.

க்ரிட்டர் ஆங்கிலத்திற்கும் பிரஞ்சுக்கும் இடையில் மொழிபெயர்ப்புக்கான மாற்றம் அடிப்படையிலான ஒழுங்குமுறையாகும். இது Centre Canadien de Recherche sur l'Infomatisation du Travail (CCRIT) கனடிய பணியிட தானியக்க ஆய்வு மையத்தால் மாண்டிரியலில் டாம் ஆய்வுத்திட்டத்துடன் இணைக்கப்பட்ட ஆய்வாளர்களால் உருவாக்கப்பட்டதாகும். இது கனடிய மாகாணங்களில் கால்நடை மற்றும் இறைச்சி வணிகச் சந்தைகளில் நிலைமையை விளக்கும் கனடிய வேளாண்மைத் துறையால் உருவாக்கப்பட்ட வார அறிக்கைகளின் குறிப்பிட்ட துணைநிலை மொழிக்கு பயன்படுத்தப்படுகின்ற பொதுவான பரிசோதனை இயந்திர மொழிபெயர்ப்பு மாதிரி ஆகும். இதன் முதன்மையான விளைவு, இலக்கண வடிவவாதமும் கணினிசார் செயல்படுத்தமும் கட்டுப்படுத்தப்படாவிட்டாலும் அகராதி அறிக்கையிகளின் சொற்றொகுதிக்கு கட்டுப்படுத்தப்பட்டுள்ளது.

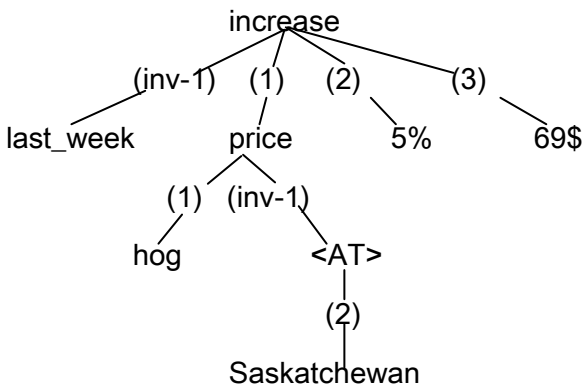
கிரைட்டர், மொழியியல் சார்ந்த இயந்திர மொழிபெயர்ப்பில் தற்காலத் தத்துவத்தின் பல வழக்கமான பண்புக்கூறுகளை எடுத்துக்காட்டும். தொடரியல் உருப்படுத்தங்கள், X-பார் மரபுகளைப் பிரதிபலிக்கும் மற்றும் சாதாரண வழியில் இடைவெளிகளையும் (gaps) சுவடுகளையும் (traces) அடையாளப்படுத்தும் மிகவும் தரமான தொடரியல் மற்றும் பொருண்மையியல் பண்புக்கூறுகளால் புலக்குறிப்புச் செய்யப்பட்ட அக அமைப்பு சார்புக் கிளை அமைப்புகளாகும் (suface structure dependency trees). வாக்கியம் 14அ விற்கு தொடரியல் கிளை அமைப்பு 14ஆ ஆகும் (Hutchins and Somers1992: 328).

14அ. Last week hog prices in Saskatchewan increased 5% at $69.

14ஆ.


S
adv_p
np
pred
<NBR:pl…>
<NBR:pl
TENSE:past,
SUBCAT:…>
v'
<NBR:pl,
TENSE:past,
SUBCAT…>
adv
n'
pp
<NBR:pl…>
n
n
p
np
v
meas_p
pp
<TYPE:proper>
<NBR:pl,
TENSE:past,
n
SUBCAT:meas_p,
<TYPE:proper>
…>
precent
p
meas_p
pice
last_week
hog
prices
in
Saskatchewan
increased
5%
at
$69

14இ



’தலைகீழான’ (‘invereted’) பங்கெடுப்பாளர் எண்கள் (argument numbers) (‘inv-1’) இரண்டு வகையான உறவுகளைச் சாத்தியமாக்கும்: பயனிலை-பங்கெடுப்பாளர் உறவுகள் மற்றும் தொடரியல் சார்பு உறவுகள். இவ்வாறு *last week* என்ற சார்பி (dependent) *increase* என்பதைப் பங்கெடுப்பாளராகக் கொண்டு ஒரு பயனிலை ஆகும் மற்றும் *prices* என்பது அருவச்சொல் <AT> என்பதற்கு ஒப்பீட்டளவில் முதல் பங்கெடுப்பாளர் இடத்தில் இருக்கின்றது (*Saskatchewan* இரண்டாவது பங்கெடுப்பாளராக இருக்கின்றது).

இடமாற்றத்திற்கு முன்னர் பொருண்மையியல் அமைப்பு பயனிலை மற்றும் பங்கெடுப்பாளர் கணுக்களின் நிலைத்தன்மைக்காகச் சரிபார்க்கப் படுகின்றது. ஒரு பொருண்மையியல் அகராதி பொருண்மையியல் ‘வகைகளை’ கணுக்களுடன் இணைக்கின்றது (எடுத்துகட்டாக MOVEMENT என்பது *increase* என்பதுடனும் MEASURE-FUNCTION என்பது *price* என்பதுடனும் INCREMENT என்பது *5%* என்பதுடனும்); மற்றும் (15) இல் உள்ளது போல் பயனிலை-பங்கெடுப்பாளர் அமைப்புகளைப் பொருண்மையியல் வகைகளுகளின் ’அமைப்புமுறைகளுக்கு’ (‘schemas’) எதிராக உறுதிசெய்கின்றது.

(15) MOVEMENT (MEASURE-FUNCTION, INCREMENT, MEASURE)

இடமாற்ற விதிகள் சொற்களுடன் தொடர்புபடுத்தப்பட்டுள்ளன. (16அ)இல் உள்ளது போல் அவை பெரும்பாலும் நேரடியானதாகும்; ஆனால் பங்கெடுப்பளர் மாற்றம் (argument conversion) (16ஆ) அல்லது கூடுதல் கலவைத்தன்மையான மாற்றம் (16இ) என்பன ‘அமைப்பு’ வேறுபாடுகளைக் கையாளலாம்.

(16அ) eat ↔ manger

(16ஆ) miss (1:X,2:Y) ↔ manquer (1:Y',2:X')

(16இ) walk (inv-1:across(2:X))↔ traverser(2:X',inv-1):$manner(2:à_pied))

க்ரிட்டரின் இலக்கணங்கள் பகுப்பாய்வுக்கும் உருவாக்கலுக்கும் ஒன்றாகும்; மற்றும் அவைகள் ப்ரொலாகில் செயல்படுத்துவதற்காக தீர்மான எச்சத்தொடர் இலக்கண வடிவாதத்தில் (Definite Clause Grammar formalism) எழுதப்பட்டுள்ளது. அவற்றிலிருந்து பகுப்பான்களும் உருவாக்கிகளும் ஆக்கப்பட்டுள்ளன; அவை முக்கியமாக விதிகள் பயன்படுத்தப்படும் நிரலில் வேறுபடுகின்றன. இதுவரை தொடரியல் பகுப்பாய்வுக்குப் பயன்படுத்தப்பட்ட இலக்கணங்கள் உருவாக்கத்திற்கும் பயன்படுத்தப்பட்டன; இடமாற்ற விதிகள் தலைகீழாக மாற்றக்கூடியது ஆகையால் க்ரிட்டர் தலைகீழாகமாற்றக் கூடிய நெறிமுறையை (reversibility methodology) எடுத்துக்காட்டுகின்றது. கிரிட்டரை உருவாக்கியவர்கள் தலைகீழாக மாற்றக்கூடிய இலக்கணங்களைக் கொண்ட ஒரே இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை அது என்று வாதிடுவர்; எடுத்துக்காட்டாக, ரோசெட்டா உருவாக்கத்தில் இல்லாத புறத் தொடரியல் கூறைக் பகுப்பாய்வில் (surface syntactic component) உட்படுத்தியுள்ளது. மாறாக, க்ரிட்டர் ஒரு விதத்தில் உண்மையில் சமச்சீரான ஒழுங்குமுறையல்ல; ஏனென்றால் பகுப்பாய்வின் ஒரு பகுதியான பொருண்மையியல் சரிபார்ப்பு உருவாகக்திற்குத் தேவை இல்லை.

3.20.8. கணினிசார் முன்னேற்றங்கள்

புலனறிவு (cognition), புலனுணர்வு (perception), கற்றல் (learning) இவற்றின் கணினிசார் மாதிரியாக்கத்தின் (computational modelling) சிறப்பான முன்னேற்றங்கள் இணைக் கணினிச்செயல்பாடு (parallel computation), நரம்பு வலைப்பின்னல்கள் (neurl neworks) மற்றும் இணைப்பாளர் மாதிரிகள் (connectionist models) என்பனவாகும். மொழி புரிந்துகொள்ளுதலின் உயர்ந்த மனச் செயல்பாடுகள், தர்க்க உய்த்தறிதல் மற்றும் நினைவகம் என்பனவற்றை மூளை போன்ற இயக்கநுட்பத்தால் மட்டுமே மாதிரியாக்கம் செய்ய இயலும் என்பது மிகப் பரவலான நம்பிக்கையாகும். அறிவை உருப்படுத்தம் செய்தல் இதே போன்று ஒன்றோடொன்று தொடர்புபடுத்தப்பட்ட 'கருத்துருக்களின்' மிக உயர்ந்த கலவைத்தன்மையான வலைப்பின்னல்களை வேண்டும். மூளையின் நரம்பு வலைப்பின்னல், கணுக்களை ஒரேசமயத்தில் அணுகவும் செயல்படுத்தவும் இயலும் என்று பரிசோதனை சான்றுகளிலிருந்து

அனுமானிக்கப்படுகின்றது. நரம்பு வலைப்பின்னல்களின் 'பரவும் செயல்படுத்தல்' ('spreading activation') மாதிரி கணினி வடிவமைப்புக்கு இணைப்பர் அணுகுமுறையால் (connectionist approach) கணினிஅடிப்படையில் மாதிரியாக்கம் செய்ய இயலும்.

இயற்கை மொழி ஆய்வுக்கு இணைப்பர் மாதிரியின் (connectionist model) சம்பந்தம் தெளிவானது. பகுப்பாய்வுக்கும் ஆக்கத்திற்கும் (உருபனியல், தொடரியல், பொருண்மையியல்) மரபு அடுக்கமைவு அணுகுமுறை (stratificational approach) மொழியியலார்களாலோ கணினி அறிவியலார்களாலோ எவ்வாறு மனிதர்கள் புரிந்துகொள்கின்றனர் மற்றும் கருத்துப்பரிமாற்றம் செய்கின்றனர் என்பதன் உளவியல் அடிப்படையிலான உண்மையான மாதிரியாக தீவிரமாக ஏற்றுக்கொள்ளப்படவில்லை. தொடரியல், பொருண்மையியல், பயன்வழியியல் செயல்முறைகளை ஒருங்கிணைக்க அடிக்கடி வெளிப்படுத்தப்படும் ஆர்வம் இணைப்பர் செயலாக்கத்தின் உள்ளார்ந்த கவர்ச்சியின் ஒப்புக்கொள்ளலாகும். எவ்வாறு உயர்ந்த ஆவலுள்ள எடிஆர் பேச்சு மொழிபெயர்ப்பு ஒழுங்குமுறை பெருமளவில் இணை கணிச்செயல்பாடு இல்லாமல் நடைமுறை படுத்த இயலும் என எண்ணுவது கடினமாகும். 'தாழ்ந்த நிலை' கண்ணோட்டத்தில் தொடரியல் அமைப்புகளின் இணை பகுப்பாய்வு கணினிச்செயல்பாட்டின் தற்போதைய வேகங்களில் குறிக்கப்பட்ட முன்னேற்றங்களை கொண்டுவரும் என்பது தெளிவாகும்.

இணை கணினிச்செயலாக்கத்தின் (parallel computation) பாதிப்பு பற்றிய ஊகம் நரம்பு வலைப்பின்னலுடன் நிரலாக்க அனுபவத்தின் பற்றாக்குறையால் சிக்கலாக்கப்பட்டுள்ளது; பெரும்பாலானவை இருக்கின்ற வரிசைக் கணிகளால் (sequential computers) உருவகப்படுத்தப்பட உள்ளது. இதுவரை சான்றுகள் இணைப்பர் வலைப்பின்னல் அதிக அளபு துல்லியத்துடன் முன்னர் காணாத வாக்கியங்களைப் பகுப்பாய்வு செய்ய 'கற்க' இயலும் என்று பரிந்துரைசெய்கின்றது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை பழைய தப்புகளிலிருந்தும் பின் திருத்திகளின் திருத்தங்களிலிருந்தும் கற்கும் கருத்து அடிக்கடி முன்வைக்கப்பட்டது. 'கற்கும்' வழக்கங்கள் செயல்படுத்தப்பட்டன; ஆனால் தவறுகள் மற்றும் திருத்தங்களின் புள்ளியியலின் அடிப்படையில் ஒழுங்குமுறையால் மாற்றங்கள் பரிந்துரைக்கப்பட்டன மற்றும் உருவாகியவர்களாலோ பயன்பாட்டாளர்களாலோ உறுதிசெய்யப்படவோ நிராகரிக்கப்படவோ செய்யப்பட்டன. உண்மையான கற்றல் ஒழுங்குமுறையில் ஒரு கலவைத்தன்மையான பின்னூட்ட இயக்கநுட்பத்தால் தானியக்கமாகத் தொடங்கப்பட்டது; மற்றும் தொடர்ச்சியாக புதிய

உள்ளீட்டால் பரிசோதிக்கப்பட்டது. இணைப்பர் மாதிரி உண்மையில் கற்கும் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் வாய்பை தந்தது.

குறைவான ஊக முன்னேற்றங்கள் பிற இயற்கை மொழி ஆய்வுகளுடன் இயந்திர மொழிபெயர்ப்பின் சாத்தியமான ஒருங்கிணைப்பு ஆகும். தகவல் மீட்பு ஒழுங்குமுறைகளுடன், அதாவது ஆர்வமுள்ள விஷயங்கள் உள்ள ஆவணங்களின் தலைப்புகள் மற்றும் சுருக்கங்களுக்காகத் தேடுவதற்குப் பயனாளிகளை தேட இயலச்செய்யும் ஒழுங்குமுறைகளுடன் வெற்றிகரமான தொடர்புகள் உள்ளன. பயனர்கள் தூரத்திலுள்ள அறியப்படாத மொழிகளில் உள்ள சுருக்கங்களைக் கொண்ட தகவல்தளங்களைத் தேடவும் சுருக்கங்களின் அல்லது முழு ஆவணங்களின் மொழிபெயர்ப்பை தமது மொழியில் வேண்டவும் இயலச்செய்கின்றது. கேள்விகள் மொழிபெயர்க்கப்படும் மற்றும் தேடலக்ள் பிறிமொழியில் நடைமுறைப் படுத்தப்படும்; அல்லது தலைப்புகள் (சுருக்கங்கள்) முன்னரே மொழிபெயர்ப்பு செய்யப்பட்டிருக்கும் (எ.கா. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையால்) மற்றும் பயனர்களின் மொழியில் தேடப்பட்டிருக்கும். தற்போது தகவல் மீட்பு ஒழுங்குமுறைகளும் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் பிரிக்கப்பட்டுள்ளன; ஆனால் எதிர்காலத்தில் ஒருங்கிணைந்த தகவல் மீட்பு ஒழுங்குமுறையை கற்பனை செய்வது சிக்கலானது அல்ல.

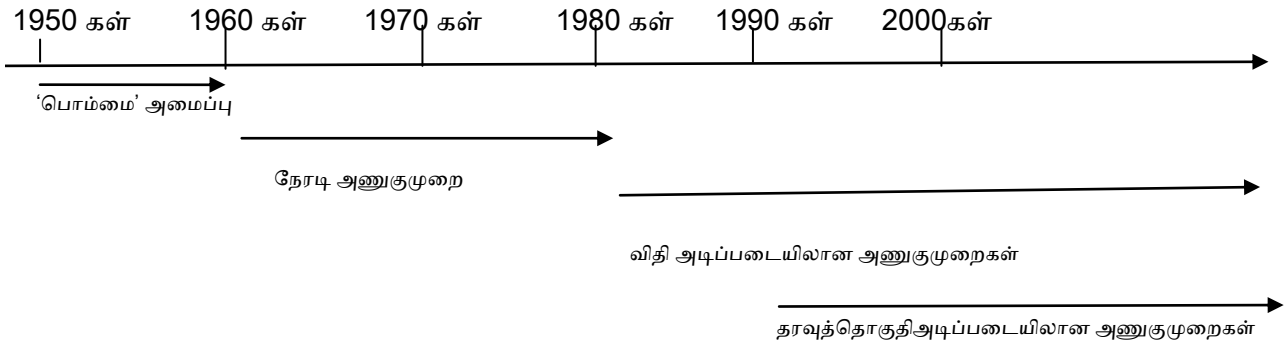
அடுத்த முன்னேற்றம் மூலமொழியில் அறிமுகமில்லாத பயனர்களுக்கு பனுவலகளின் சுருக்கங்கள் அல்லது சுருக்கத்திரட்டுகளின் தானியக்க உருவாக்கம் ஆகும். வெளிநாட்டு மொழி ஆணவங்களின் சுருக்கங்கள் பெரும்பாலான நிர்வாகிகளுக்கும் வணிகர்களுக்கும் அறிவியல் அறிஞர்களுக்கும் முழுப் பனுவல்களின் முரட்டுத்தனமான மொழிபெயர்ப்புகளைவிட கவர்ச்சியானதாகும். இருப்பினும், கட்டுப்படுத்தப்பட்ட பொருண்மைக் களங்களில் சுருக்க உரையாக்கம் மீதான சிறிய அளவிலான பரிசோதனைகளிலிருந்து இதன் சிக்கல்கள் இயந்திர மொழிபெயர்ப்புக்குச் சமமானவை என்று தெரியவருகின்றது.

3.21. சுருக்கம்

இயந்திர மொழிபெயர்ப்பு ஐந்து பதின்ம ஆண்டுகளுக்கும் மேலானது. இருப்பினும் ‘இயந்திர மொழிபெயர்ப்பு’ கிட்டத்தட்ட நானூறு ஆண்டுகளுக்கு முன்பு தோன்றியது. 1629இல் டெஸ்கார்ட்ஸ் (Descartes) ஒரு மொழியைக் குறியீடுகள் மற்றும் வெவ்வேறு மொழிகளின் சொற்களால் குறிப்பிடப்படலாம் என்றும் சமமான பொருள் உள்ள மொழிகள் ஒரே குறியீட்டைப்

பகிரக்கூடும் என்ற கருத்தை முதலில் முன்வைத்தவர் எனக் கருதலாம் (Pugh 1992: 15). இது தனிப்பயன்-வடிவமைக்கப்பட்ட இயந்திரம் (custom-designed machine) இயற்கை மொழியின் குறியிடப்பட்ட விதிகளை கையாள முடியும் என்ற கருதுகோளின் அடிப்படையில் அமைந்தது ஆகும். இது இருந்திருக்கலாம் நவீன இயந்திர தொழில்நுட்பத்தின் முன்னோடி (Wilss 1999: 140). இயந்திர தொழில்நுட்பத்தின் ஆரம்ப காலத்தில் 'தானியங்கி மொழிபெயர்ப்பு' 'automatic translation' அல்லது 'இயந்திர மொழிபெயர்ப்பு' ('mechanical translation') என்ற பொதுவான சொற்கள் பயன்படுத்தப்பட்டன (Tong 1994: 4,731). எனினும், அது இரண்டாம் உலகப் போருக்குப் பின்னர்தான் (1939-45) அந்த நேரத்தின் சமீபத்திய கண்டுபிடிப்பான சேமிக்கப்பட்ட-நிரல் கணினிகளைப் (stored-program computers) பயன்படுத்தி மொழி மொழிபெயர்ப்பு செய்யும் சாத்தியக்கூறுகள் ஆராயப்பட்டன. (Somers 2003b: 4) படம் 3.1 நேரத்தின் தொடர்ச்சியான தோராயத்தைக் காட்டுகிறது இயந்திர மொழிபெயர்ப்பு அமைப்பு வளர்ச்சியில் பயன்படுத்தப்படும் அணுகுமுறைகள்

கீழ் காணும் படம் இருபதாம் நூற்றாண்டின் இரண்டாம் பாதியிலிருந்து காலங்களின் அடிப்படையில் எந்த ஒரு குறிப்பிட்ட அணுகுமுறை அதன் உச்சத்தில் இருந்தது என்பதை அறிய ஒரு வழிகாட்டியாகச் செயல்படும்.



இயல் 4

இந்தியாவில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி

4.1.அறிமுகம்

ஆரம்பக்கட்டத்தில் இந்தியாவில் இயந்திர மொழிபெயர்ப்புகள் இதற்கான சாத்தியத்தை வெளிப்படுத்தும் பொம்மை ஒழுங்குமுறைகளாக (toy systems) இருந்தன. 1978 வரை இயந்திர மொழிபெயர்ப்பிற்கான எந்த முயற்சியும் எடுக்கப்படவில்லை. 1978-இல் ஐ.ஐ.டி.கான்பூரில் உள்ள மின்னியல் தொழில் நுட்பத் துறையின் முயற்சியால் மொழியியல் அடிப்படையாலன தகவல் ஒழுங்குமுறைகள் பற்றிய தேசியக் கருத்தரங்கு (National Symbosium on Linguistic Based Information System) முதன் முதலில் நடைபெற்றது. இக்கருத்தரங்கு ஆராய்ச்சியாளர்களிடையேயும் அரசாங்கத்திலும் தொழில்நுட்பத் துறைகளிலும் இயந்திர மொழிபெயர்ப்பு குறித்த விழிப்புணர்ச்சியை ஏற்படுத்தியது. இருப்பினும் இயந்திர மொழிபெயர்ப்பு 80-களின் ஆரம்பகட்டத்தில் இந்திய மொழிகளுக்குச் செல்லாய்விகளும் பிற கணிபொறி வசதிகளும் வரத் தொடங்கிய பின்னர்தான் தொடங்கப்பட்டது. ஐ.ஐ.டி.கான்பூர் GIST என்ற தொழில் நுட்பத்தின் உருவாக்கத்திற்கு முயன்றது. இது இயந்திர மொழிபெயர்ப்பிற்கு விழிப்புணர்ச்சி ஏற்படுத்துவதுடன். இதில் ஆய்வு மற்றும் வளர்ச்சி குறித்த பல திட்டங்களை ஊக்குவித்தது. இந்தியாவில் பல மையங்களில் பலவிதமான முயற்சிகள் மேற்கொள்ளப்பட்டாலும் மூன்றும் அணுகுமுறைகளைப் பட்டியலிட இயலும்.

இடமொழி அணுகுமுறை (Interlingual Approach)

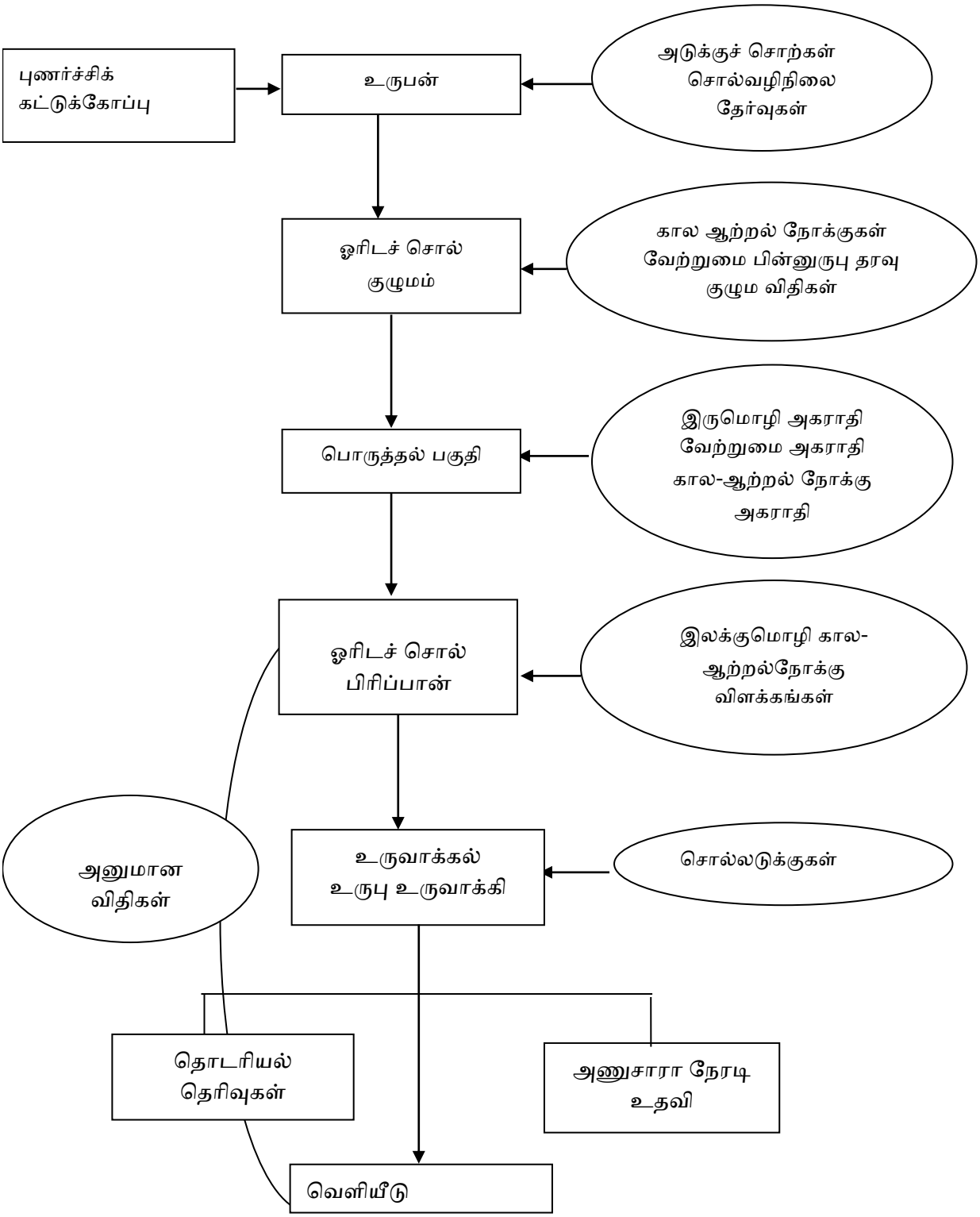
நேரடி சொல்சார் மாற்றல் அணுகுமுறை (Direct lexical Transfer Approach)

கணிப்பொறி உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிகள் (Machine Aided Translation Effort)

4.2.அனுசாரக் இயந்திர உதவியிலான மொழிபெயர்ப்பு ஒழுங்குமுறை

சமஸ்கிருதத்தை இடைப்பட்ட அடிப்படை மொழியாகக் கொண்டு இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்ப்பு செய்யும் இடைமொழி ஆய்வு அணுகுமுறை சின்ஹா என்பவரால் 1984-இல் முயற்சிக்கப்பட்டு 1989-இல் விரிவாக்கப்பட்டது. இது மூல மொழியின் ஆய்வுக்குக் காரக அடிப்படையிலான அக உருப்படுத்தத்தை முன்மொழிந்தது. 1986-1988-களில் எல்லைகுட்பட்ட சொற்களைக் கொண்ட மிக எளிய வாக்கியங்களை இந்தியிலிருந்து

தெலுங்கிற்கு மொழிபெயர்க்கும் மூலமுன்மாதிரி (prototype) சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் உருவாக்கப்பட்டது (Akshar Bharti et al 1997) . இது காரக அடிப்படையிலான கருத்துவரைபடத்தை அக உருப்படுத்திற்கும் பொருண்மை மயக்கங்களுக்கும் பயன்படுத்தியது. ஒரு தனிநிலை தொடரியல் இருப்பானில் உள்ள மிகக்கூடுதலாகத் தொடர்புள்ள சொற்கள் ஒரு சொல் குழுவாகப் பயன்படுத்தப்பட்டது. மூல மொழியில் உள்ள உரை உருவாக்கி இச்சொற் குழுமங்களிலிருந்து பொருத்தத்தைச் சரியான உருபனியல் உருவாக்கத்திற்குப் பயன்படுத்தியது.

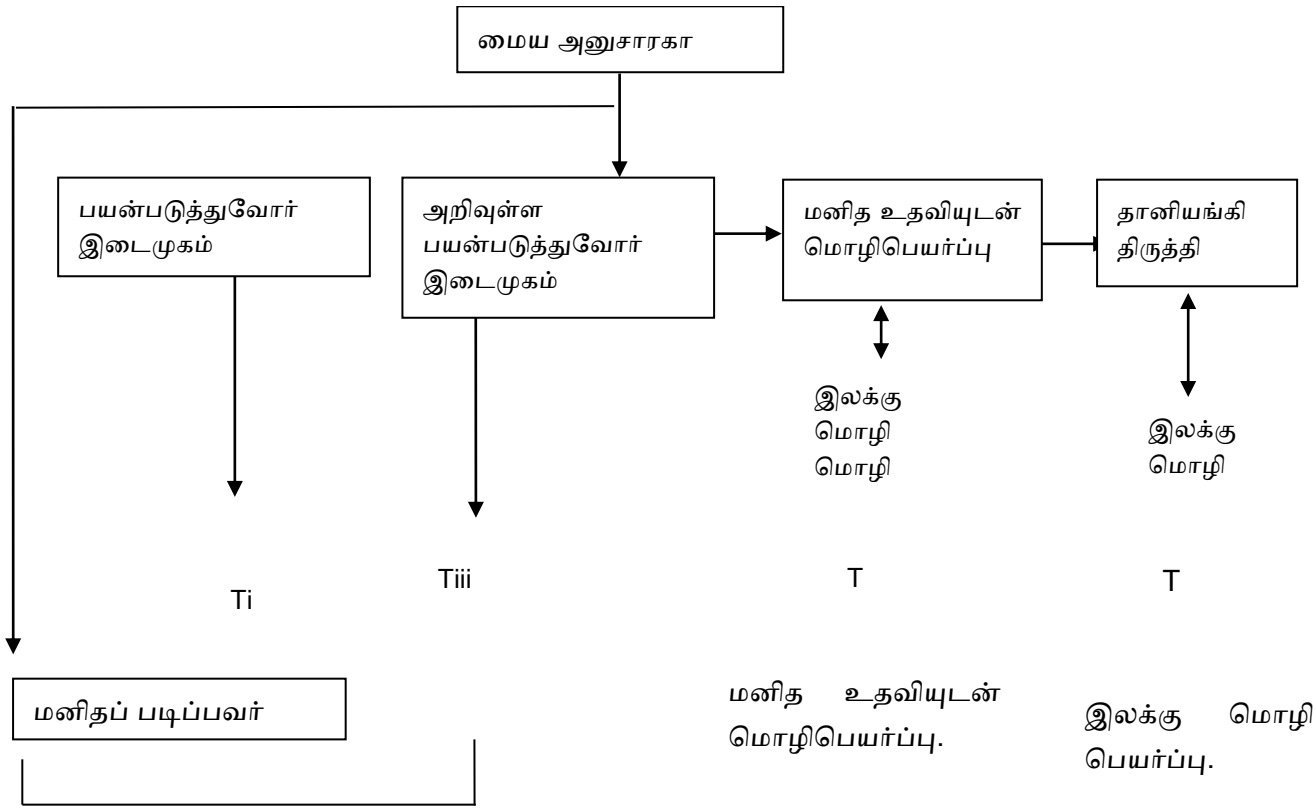

புணர்ச்சிக் கட்டுக்கோப்பு
உருபன்
அடுக்குச் சொற்கள்
சொல்வழிநிலை
தேர்வுகள்
ஓரிடச் சொல் குழுமம்
கால ஆற்றல் நோக்குகள்
வேற்றுமை பின்னுருபு தரவு
குழும விதிகள்
பொருத்தல் பகுதி
இருமொழி அகராதி
வேற்றுமை அகராதி
கால-ஆற்றல் நோக்கு அகராதி
ஓரிடச் சொல் பிரிப்பான்
இலக்குமொழி கால-ஆற்றல்நோக்கு விளக்கங்கள்
அனுமான விதிகள்
உருவாக்கல் உருபு உருவாக்கி
சொல்லடுக்குகள்
தொடரியல் தெரிவுகள்
அணுசாரா நேரடி உதவி
வெளியீடு

இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயரிப்பு அதற்குப்பின் முயற்சிக்கப்படவில்லை. இவ்வணுகுமுறை இந்திய மொழிகளின் பன்மைத்தன்மைக்குக் கூடுதல் அனுகூலமாய் இருந்தாலும் அக உருப்படுத்ததிலிருந்து புற அமைப்பை பெறுவது கடினமான செயல்பாடாக அமைந்து குறையுடையதாய் இருந்தது. இவ்வணுகுமுறை மூலமொழியின் சீரான புரிதலை முன்கருதலாகக் கொண்டதால் இன்றைய ஆய்வுநிலையில் ஏற்றதாக அமையவில்லை. மேலும் அமைப்பு அடிப்படையில் அண்மைப்படுகின்ற இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்பின் போது நேரிடையான சொல் பதிலீடு செய்வது எளிமையாக அமைந்தது. இதன் காரணமாக இவர்களால் அனுசாரகா என்ற அணுகுமுறை உருவாக்கப்பட்டது (Durgesh et al., 2000; Sudip et al., 2005),.

இந்திய மொழிகளுக்கிடையே உள்ள அமைப்பு ஒற்றுமை காரணமாக மூலமொழியிலிருந்து இலக்கு மொழிக்குப் பல பொருண்மை மயக்கங்களைக் கொண்டு செல்ல இயலும் என்ற காரணத்தால் இந்த இடைமொழி அணுகுமுறை இந்திய மொழிகளுக்கிடையே மொழிபெயர்ப்பிற்கு நல்லதொரு அணுகுமுறையாக அமையும். இருப்பினும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து ஆங்கில மொழிகளுக்கும் மொழிபெயர்க்க இடைமொழி அணுகுமுறை பொருத்தமான அணுகுமுறை அல்ல.

இயந்திர உதவியிலான மொழிபெயர்ப்பு ஒழுங்குமுறையான அனுசாரக (ANUSAARAKA) சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் மூலமொழியிலிருந்து இலக்கு மொழிக்கு நேரடி சொல்சார் மாற்ற அணுகுமுறையாக உருவாக்கப்பட்டது. இங்கு மூல மொழிக்கும் இலக்கு மொழிக்கும் இடையில் உள்ள வாக்கிய அமைப்பின் பொதுமைப் பண்புகள், இரண்டும் இந்திய மொழிகள் என்ற காரணத்தால் கருதப்பட்டு முழு அளவிலும் பயன்படுத்தப்பட்டுள்ளது. இவ்வணுகுமுறையின் படி ஒரு தனிநிலை தொடரியல் உறுப்பாக நடத்தை செய்யும் மூலமொழியிலுள்ள சொற்கள் குழுமப்பட்டு மூலமொழிக்குப் பொருந்தும்படி அது சொற்களால் பதிலீடு செய்யப்படும் பின்னுருபை நிர்ணயிக்க சில விதிகள் உருவாக்கப்பட்டன. சொற்களின் நிரல் வருகை பெரும்பாலும் தக்க வைக்கப்பட்டு சில சமயங்களில் புதிய இணைப்பான்களின் பயன்பாட்டால் நிறைவு செய்யப்படும். சில சமயங்களில் இம்முறை வெளியீட்டில் தொடரியல் தவறுகளுக்கு வழிவக்கும்; இது பற்றி இதை உருவாக்கியவர்கள் கவலைப்படவில்லை. இருப்பினும் குறைந்த அளவு பொருள் பெரும்பாலும் வெளிப்படுத்தப்படும். மேலும் அனுசாரக்காவின் பாகங்கள் அல்லது உறுப்புத் தொகுதிகள் எந்த இயந்திர மொழிபெயர்ப்பிற்கும் அல்லது இயந்திர உதவியுடன் செய்யப்படும் மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கும் பகுதியாக அமையும்.

இவ்வாறு அனுசாரகாவின் வெளியீடு இடைப்பட்ட ஒரு நிலைக்கு முக்கிய வெளியீடாகும் மற்றும் அனுசாரகா போன்ற வெளியீடு ஒரு இரண்டாம் நிலை விளைவாகும் (by product). இது சீரற்ற மொழிபெயர்ப்பிற்கு கொண்டுசெல்லாததால் மொழிகளின் இணையின் இயந்திர மொழிபெயர்ப்பிற்கும் பயனுள்ள உபாயமாகும். இதை உருவாக்கியவர்கள் இதை ஒரு மொழி அணுகி (language acessor) என்று அழைகின்றனர்; இதிலிருந்து முழு நன்மை அடைய பயன்படுத்துபவர்களிடம் சில பயிற்சியை எதிர்பார்க்கின்றனர். தொடக்கத்தில் கன்னடா, இந்தி மொழிகளுக்கு ஒரு அனுசாரக் ஒழுங்குமுறை அமைக்கப்பட்டு காட்டப்படது. இதன் பின்னர் தெலுங்கு-இந்தி, பஞ்சாபி-இந்தி, மராட்டி-இந்தி, சமஸ்கிருதம்-இந்தி, தமிழ்-இந்தி ஆகிய மொழி இணைகளுக்கு அனுசாரக் ஒழுங்குமுறைகள் உருவாக்கப்பட்டு பார்வையாளர்களுக்கு எடுத்துக்காட்டப்பட்டது.



அனுசாரகா வெளியீட்டில் பல மட்டங்கள்

(http:/iiit.net/ltrc/Anusaaraka/anu_home.html).

4.3. சிவ மற்றும் சக்தி ஒழுங்குமுறை (SHIVA and SHAKTI MT System)

யு.எஸ்.ஏ.இல் உள்ள கார்னெகி மெல்லொன் பல்கலைகழகமும் (Carnegie Mellon University) (Sudip et al., 2005; Dwivedi et al., 2010), பெங்களூரில் உள்ள இந்திய அறிவியல் நிறுவனமும் (Indian Institute of Science, Bangalore), ஹைதராபாத்திலுள்ள உலகத் தகவல் தொழில்நுட்ப நிறுவனமும் (International Institute of Information Technology, Hyderabad) இணைந்து ஆங்கிலத்திலிருந்து இந்திக்கு மொழிபெயர்ப்பதற்காக சிவ (SHIVA) மற்றும் சக்தி (SHAKTI) என்ற இரண்டு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்கியது. இவ்விரு மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் பரிசோதனைக்காகவும் வெள்ளோட்டமாகவும் பயன்படுத்துபவரின் கருத்தறியவும் வெளியிடப்பட்டன. சிவ என்ற மொழிபெயர்பு ஒழுங்குமுறை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையாகும். (சிவ ஒழுங்குமுறையை பின்வரும் வலை தளத்தில் பெறலாம்: (http://ebmt.serc.iisc.ernet.in/mt/login.html மற்றும் சக்தி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை பின்வரும் வலை தளத்தில் பெறலாம்: http://shakti.iiit.net). 2006 ஏப்ரலில் சக்தியின் 0.81 மாதிரி வெளியிடப்பட்டது. இது இந்தி, மராத்தி, தெலுங்கும் என்ற மூன்று இலக்கு மொழிகளுக்குச் செயல்புரியும். பயன்பாட்டாளர் ஆங்கில வாக்கியங்களை மொழிபெயர்க்க முதலில் இந்த இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையுடன் இணைய வேண்டும். ஒழுங்குமுறை பட்டியலிருந்து எழுத்துருவை தெரிந்தெடுக்கக் கேட்கும்; மொழிபெயர்க்கப்பட்ட வெளியீட்டை தெரிந்தெடுக்கப்பட்ட எழுத்துருவில் காட்டும். பயன்பாட்டாளர் மொழிபெயர்ப்பு ஒழுங்குமுறையால் உருவாக்கப்பட்ட மொழிபெயர்ப்பில் திருப்தி பெறவில்லை என்றால் ஒழுங்குமுறை பயன்பாட்டாளரிலிருந்து மூலமொழியின் பல்வேறு கூறுகளின் (சொல், தொடர், வாக்கிய நிலைகளில்) பொருள்களைக் கேட்டு உள்ளீட்டுப் பனுவலை மீள் மொழிபெயர்ப்பு செய்யும்.

4.4. **ஆங்கிலபாரதி** (*ANGLABHARATI*)

ஐ.ஐ.டி.கான்பூரில் மேற்கொள்ளப்பட்ட ஆங்கில பாரதி (*ANGLABHARATI*) திட்டத்தை நிறுவுதில் நான்கு முக்கிய கருதல்கள் இருந்தன. முதலாவது அன்றைய காலகட்டத்தின் நிலையில் சீரான இயந்திர மொழி பெயர்ப்பு சாத்தியமில்லை. எனவே இயந்திர உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிக்கப்பட வேண்டும்; இங்கு இயந்திரம் பெரும்பாலான செயல்களைச்

செய்கின்றது. கடினமான வேலைகளள் முந்தைய மற்றும் பிந்தைய திருத்தங்களால் (Pre and Post editings) கையாளப்படுகின்றது. இதன்படி ஒவ்வொரு மூலமொழி வாக்கியத்ததையும் தனியாகப் பரிசோதிப்பது மற்றும் முற்கட்டு மற்றும் (anaphora) வாக்கியங்களுக்கிடையிலான குறிப்புகள் (Intersentential References) போன்ற சிக்கல்களை மனிதத் திருத்துனருக்கு விட்டுவிடுவது என்பன எளிமைக்கு வழி வகுக்கிறது. இரண்டாவது இயந்திர மொழி பெயர்ப்பு, பயனப்படுத்துவோர் மற்றும் பயன்பாட்டு அடிப்படையில் இருக்கவேண்டும் என்ற உணர்வு, ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு உள்ள மொழிபெயர்ப்பை மேற்கொள்ளுவது இந்திய மொழிகளுக்கு இடையிலான இயந்திர மொழிபெயர்ப்பை மேற்கொள்வதைக் காட்டிலும் சிறந்தது என்ற தீர்மானத்திற்கு வழிவகுத்தது. மூன்றாவது ஒரு குறிப்பிட்ட பொருண்மைக்களத்தை அடிப்படையாகக் கொண்ட ஒழுங்குமுறைகள் (domain specitic systems) திறந்த ஒழுங்குமுறைகளைக் காட்டிலும் (open ended systems) உருவாக்குவதற்கு எளியது மற்றும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பு பயன்படுத்துவோர்க்கு சிக்கனமானதாகவும் இருக்கிறது. இறுதியாக வேறுப்பட்ட மொழிகள் ஒற்றுமையுள்ள மொழிகளைக் காட்டிலும் ஆய்வு அடிப்படையில் கூடுதல் அறை கூவல்களை எதிர்கொள்ளும்.

சில முக்கியத்திட்டக் கருத்தல்கள் 90 விழுக்காடு வேலையை இயந்திரத்தாலும் வேலையை மனித முன் திருத்தத்தாலும் (post editing) மேற்கொள்ளும் படிக்கு மொழிப் பெயர்ப்பிற்கு ஒரு பயன்பாட்டு கருவியைத் தருவது என்ற நோக்கம் அடிப்படையில் அமைந்தது, இந்த அமைப்பொழுங்கு படிப்படியாகச் சிக்கலான சூழல்களைக் கையாளும்படிக்கு வளர இயலும், பொத்துமான உரை உருவாக்கும் பகுதிகளின் (text generator modules) இணைப்பால் ஆங்கிலத்திலிருந்து பெரும்பாலான இந்திய மொழிகளுக்கு மொழிப்பெயர்ப்பு ஒரு சீரான இயங்கு முறையைக் கொண்டிருந்தது. மேலும் அதன் பயன்பாட்டிற்கும் விரிவாகத்திற்கும் வசதி செய்ய மனிதனால் இயக்கப்பட்ட மனிதன் - இயந்திர இடைமுகம் கொண்டிருந்தது.

இது ஒரு குழும இந்திய மொழிகளுக்கு (target languages) பயன்படுமாறு போலி இலக்குமொழியை உருவாக்கும் (Pseudo target language) சூழல் வரையறையற்ற இலக்கணம் போன்ற அமைப்புடன் கூடிய அமைப்பொழுங்கை இலக்காகக் கொண்ட விதி அடிப்படையிலான ஒழுங்குமுறையாகும். பெருந்தரவு ஆய்வின் மூலம் கிடைக்கப்பெற்ற ஒரு குழும விதிகள் போலி மூலமொழிக்கு நகர்வு விதிகளைப் பயன்படுத்தும் சாத்தியமான உறுப்புகளைக் கண்டு கொள்வதற்குப் பயன்படுத்தப்படுகிறது. போலி இலக்குமொழியைப் பயன்படுத்தும் கருத்து,

இடைமொழி அணுகுமுறையைப் பயன்படுத்துவதற்குச் சமமான நன்மையைப் பெற அமைப்புசார் ஒற்றுமையை முதன்மையாகப் பயன்படுத்துவதாகும். போலி இலக்குமொழி, இடைமொழி அணுகுமுறையில் பயன்படுத்தப்படும் இடைப்பட்ட மொழி அல்ல இங்கு பொருண்மை உருப்படுத்தத்தை உருவாக்க எந்த முயற்சியும் மேற்கொள்ளப்படவில்லை.

இந்திய மொழிகள் சுதந்திர சொல்குழும நிரலைக் கொண்டு வினையில் முடியும் பண்புடையனவாகும். அமைப்பு ஒற்றுமையின் வழி இந்திய மொழிகளை அவற்றின் மூலம் அடிப்படையில் நான்கு பெரிய குழுமங்களாக வகைப்படுத்தலாம்:

1. இந்திய – ஆரிய மொழிக்குடும்பம் (Indo-Aryan family) (இந்தி, வங்காளம் அஸ்ஸாமிஸ், பஞ்சாபி, மராத்தி, ஒரியா குஜாத்தி போன்றன)

2. திராவிட மொழிக் குடும்பம் (Dravidian Family) (தமிழ், தெலுங்கு, கன்னடம், மலையாளம் போன்ற)

3. ஆஸ்ட்ரோ – ஆசிய மொழிக் குடும்பம் (Astro-Asian family),

4. திபெத்திய பர்மிய மொழிக் குடும்பம் (Tibato-Burman Family).

ஒவ்வொரு குழுமங்களுக்குள்ளும் மொழிகள் உயர்ந்த அளவு அமைப்புப் பொருத்தத்தைக் காட்டுக்கின்றன. ஆங்கில பாரதி திட்டம் இந்த ஒற்றுமையைக் கூடுதல் அளவு இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கப்பயன்படுத்துகிறது.

இலக்குமொழியில் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கு வேண்டி பல பொருண்மை அடையாளங்கள் பயன்படுத்தப்படுகின்றன. தீர்க்கப்படாத பொருண்மை மயக்கங்களின் மாற்றுப் பொருள்கள் போலி இலக்குமொழியில் தக்க வைக்கப்படுகிறது. ஒவ்வொரு இலக்குமொழிக்கும் உரை உருவாக்கப் பகுதி போலி இலக்குமொழியை இலக்குமொழிக்கு மாற்றுகிறது. ஒவ்வொரு இலக்குமொழிக்கும் திருத்துவான் (corrector) பயன்படுத்தப்படுகிறது. இறுதியாக மனித இயக்கத்தால் தூண்டப்பட்டபின் திருத்தும் தொகுதி (Post editing package) இறுதி திருத்தங்களைச் செய்வத்றகுப் பயன்படுத்தப்படுகிறது. பின்திருத்தி இலக்குமொழியை மட்டும் அறிந்தால் போதுமானது.

ஆங்கில பாரதி திட்டம் சின்ஹா (Sinha, 1993) என்பவரால் 1991-ல் உருவாக்கப்பட்டது. இத்திட்டத்தில் ஆங்கிலத்திலிருந்து இந்தி மற்றும் தெலுங்கு மொழிகளுக்குச் செயல்படக்கூடிய மூலமுன்மாதிரி மொழிபெயர்ப்பு (Functional Prototype) ஒழுங்குமுறை உருவாக்கப்பட்டது.

பின்வரும் திட்டவரைபடம் இந்த மொழி பெயர்ப்பு ஒழுங்குமுறையை விளக்கும். பின்வருவன ஆங்கில பாரதி மொழிபெயர்ப்புத் திட்டத்தின் முக்கியக்கூறுகளாகும்.

1. விதி அடிப்படை

விதி அடிப்படை (Rule base) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு வாக்கியங்களின் அமைப்புகளைப் பொருத்துவதற்கான விதிகளைக் கொண்டிருக்கின்றது. இந்த ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு அமைப்பு ஒழுங்கு மாற்றங்களின் தரவு மையம் மொழிபெயர்களுக்கு அமைப்பு ஒழுங்கு மாற்றங்கள் தரவு மையம் மொழிபெயர்க்கவேண்டிய வாக்கியத்திற்காகக் கிளை அமைப்பைப் பெறும் வேலையைப் புறக்கணித்து புறக்கிளைகளிலிருந்து புறக்கிளை மாற்றத்தை செய்யும் வேலைக்குப் பணிக்கப்படுகிறது. வாக்கியங்கள் பொருண்மையைக் கண்டு பிடிக்க புற அமைப்பொழுங்கைப் பயன்படுத்தும் கருத்து, மொழியியலில் மிகப் பழமையானதாகும். இந்த அணுகுமுறை எளிதானது என்றாலும் இத்திட்டம் ஒரு மொழியின் புற அமைப்பு ஒழுங்களின் தனிப்பட்ட தன்மைகளைக் கண்டுபிடிக்கச் செயலாக்கமுடையதாகும். ஆங்கில பாரதியில் பயன்படுத்தப்படும் அணுகுமுறை சாம்ஸகியின் தொடரமைப்பு இலக்கணத்திலிருந்தும் சொல்சார் செயல்பாடுசார் இலக்கணத்தின் (Lexical functional Grammer) அமைப்பிலிருந்தும் (C-structure) அதிகமாக எடுத்தாண்டுள்ளது. விதி அடிப்படை (rule base) என்று கூறப்படும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கான அமைப்பு மாற்ற விதிகளின் தரவு மூலம் ஆங்கில பாரதி ஒழுங்குமுறையில் மையப் பகுதியாக அமைகிறது. இது ஆங்கிலத்திலிருந்து மொழிபெயர்கும் போது தொடரியலில் ஏற்படும் முக்கிய மாற்றங்களைக் கவனித்துக்கொள்கிறது. முன்னர் கூறியப்படி இந்திய மொழிகளுக்குப் பொது விதி அடிப்படையை உருவாக்கி ஆங்கில பாரதி ஆங்கிலத்திலிருந்து மொழிபெயர்க்கும் போது சிறந்த நன்மையை வெளிப்படுத்துகிறது.

2. அர்த்த மயக்க நீக்கி

அர்த்த மயக்கநீக்கி (Sense disambiguator) மூல மொழியின் ஒவ்வொரு சொல்லுக்கும் சரியான அர்த்ததை தேர்ந்தெடுக்கப் பொறுப்புள்ளதாகும். இங்கு அர்த்த மயக்கங்க நிக்கம் மூல மொழியின் உரையில் மட்டுமே நிகழ்த்தப் பெறுகின்றது. ஆங்கில பாரதியில் பயன்படுத்தப்படும் இவ்வணுகுமுறையை விதியால் விதி-பொருள்கோள் (Rule-by-rule Semantic Interpretation) எனக் கூறலாம். ஒரு தொடரியல் விதி பயன்படுத்தப்படும் ஒவ்வொரு நேரத்திலும் பொருள்கோளி (Semantic Intepreter) பயன்படுத்தப்படுகின்றது.

3. இலக்கு உரை உருவாக்கி

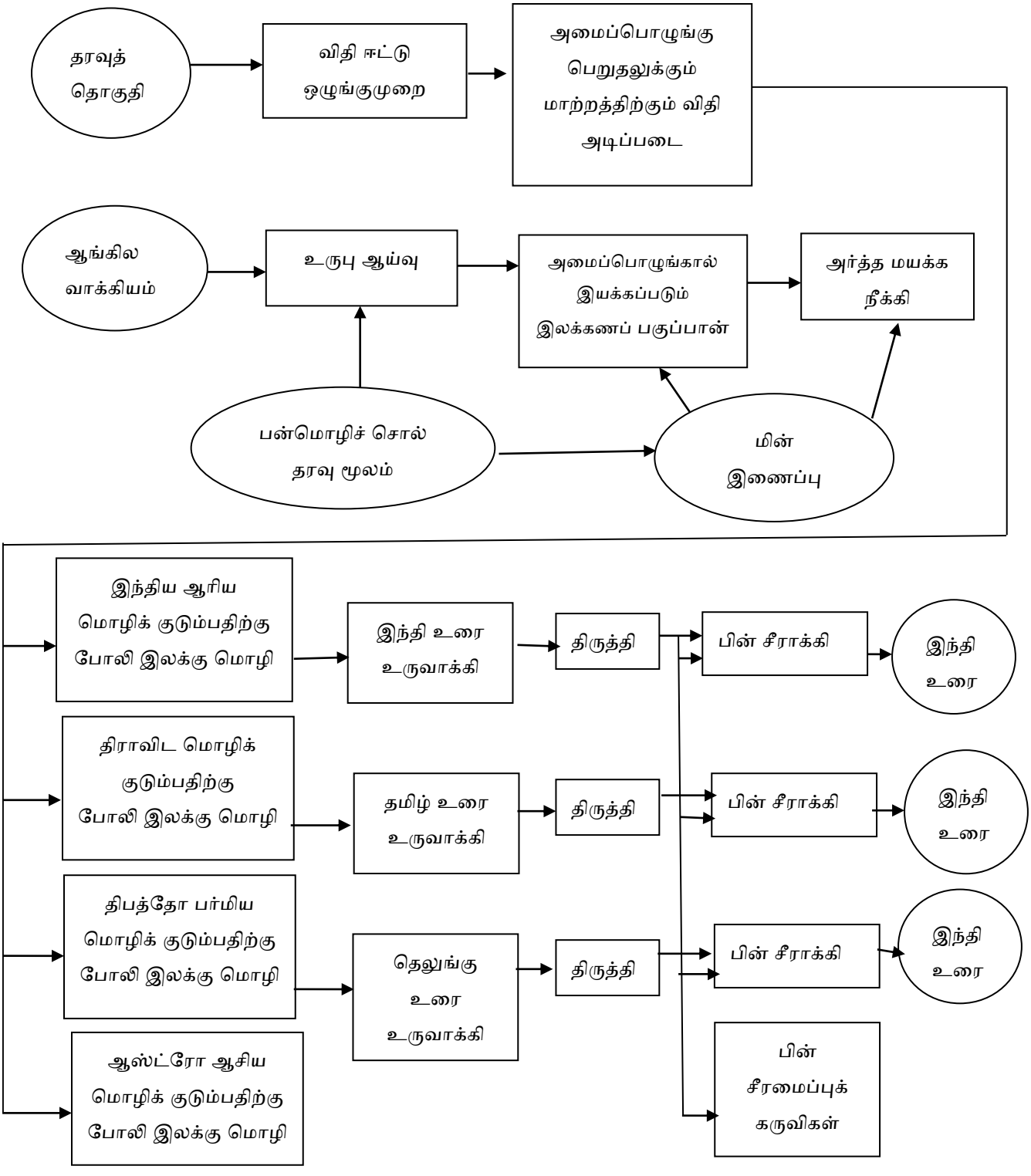
இலக்கு உரை உருவாக்கி (Target text generator) மொழிபெயர்ப்பு ஒழுங்குமுறையின் இறுதியாக அமைகின்றது. இதன் செயல்பாடு பொறுத்தமுறும் மூல மொழிகளிலிருந்து மொழிபெயர்க்கப்பட்ட வெளியீட்டை உருவாக்குவது ஆகும். இது ஆங்கில பாரதியின் முந்தைய நிலைகளின் உருவாக்கப்பட்ட இடைப்பட்ட வடிவத்தை உள்ளீடாக ஏற்றுக் கொள்கின்றது. இவ்வேலையை இயற்கை மொழி உருவாக்கம் (Natural Language Generation) என்று அழைக்கப்படும் வேலையிலிருந்து வேறுபட்டவையாகும்; இவ்வர்த்ததில் பிந்தையது எனக் கூறவேண்டும் என்பதுடன் திட்ட நிலை (Strategic level) எவ்வாறு கூற வேண்டும் என்பதையும் நடவடிக்கை நிலை (Tactical Level) தீர்மானிக்க வேண்டும். ஒரு விதி அடிப்படை மற்றும் அர்த்த மயக்க நீக்கி இவற்றைப் பயன்படுத்தும் வேறுபட்ட உரை உருவாக்கிகளைக் கொண்டு பல இலக்கு மொழிகளுக்கும் பொது இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை அமைப்பு கிடைக்கப்பெறுகின்றது.

4. பன்மொழி அகராதி

பன்மொழி அகராதி (Multilingual Dictionary) தொடரியல் வகைப்பாடுகள், சாத்தியமான அர்த்தங்கள், அர்த்தங்களின் மயக்கத்தை நீக்கும் முக்கிய கூறுகள் (Keys) இவற்றை உள்ளடக்கிய ஆங்கிலச் சொற்களுக்கும் இணையாக மூலமொழிச் சொற்களுக்குப் பல விளக்கங்களைக் கொண்டிருக்கும்.

5. விதி அடிப்படை ஈட்டி

விதி அடிப்படை ஈட்டி (rule base acquirer) மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு விதி அடிப்படையை உருவாக்குகிறது.


தரவுத் தொகுதி
விதி ஈட்டு ஒழுங்குமுறை
அமைப்பொழுங்கு பெறுதலுக்கும் மாற்றத்திற்கும் விதி அடிப்படை
ஆங்கில வாக்கியம்
உருபு ஆய்வு
அமைப்பொழுங்கால் இயக்கப்படும் இலக்கணப் பகுப்பான்
அர்த்த மயக்க நீக்கி
பன்மொழிச் சொல் தரவு மூலம்
மின் இணைப்பு
இந்திய ஆரிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
இந்தி உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
திராவிட மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
தமிழ் உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
திபத்தோ பர்மிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
தெலுங்கு உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
ஆஸ்ட்ரோ ஆசிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
பின் சீரமைப்புக் கருவிகள்

4.5.அனுபாரதி

1995-இல் சின்ஹா அனுபாரதி வழிமுறையை உருவாக்கினார். இது எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு உபாயத்தைப் பின்பற்றியது. இந்த வழிமுறை இந்தியிலிருந்து ஆங்கிலத்திற்கு மொழிபெயர்க்கப் பயன்படுத்தப்பட்டது. அனுபாரதியின் அணுகுமுறை ஒற்றுமையுள்ள மொழிகளுக்கிடையில், எடுத்துக்காட்டாக இந்திய மொழிகளுக்கு இடையில் இயந்திர மொழிபெயர்ப்பு செய்யத் திறமையாகச் செயல்பட்டது. அம்மாதிரியான நேர்வுகளில் சொல்வரிசை ஒன்றாக இருக்கும் மற்றும் பொருத்தங்களை நிறுவதற்கு சுட்டிக்காட்டிகள் தேவையில்லை.

ஆங்கிலபாரதி, அனுபாரதி ஆகிய இரண்டு இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளும் அதன் தொடக்க கருத்துருவாக்கத்திலிருந்து கணிசமான மாற்றங்கள் அடைந்தன. 2004-இல் மொழிபெயப்பு ஒழுங்குமுறையின் முன்னேற்றத்தின் இரண்டாவது கட்டம் வெளியிடப்பட்டது. இது முந்தைய கட்டமைப்பின் குறைகளை நிவர்த்தி செய்தது. இவை ஆங்கிலபாரதி II என்றும் அனுபாரதி II என்றும் பெயரிடப்பட்டன. இவ்விரு இயந்திரபெயர்ப்பு ஒழுங்குமுறைகளும் வேறுபட்ட முன்னுதாரணங்களின் கலப்பாக்கத்தின் மாறுபடும் அளபுகளில் கலப்பாக்கம் செய்யப்பட்டுள்ளன.

4.6.ஆங்கிலபாரதி II (2004)

ஆங்கிலபாரதி-II (Sinha, 2004) முந்தைய மொழிபெயர்ப்பு ஒழுங்குமுறையின் கட்டமைப்பின் பல குறைபாடுகளை நீக்கியுள்ளது. இது கலப்பிற்காகப் பொது எடுத்துக்காட்டு அடிப்படையையும் (Generalized Eample-base (GEB)) மூல எடுத்துக்காட்டு அடிப்படையையும் (Raw Example-base (REB)) பயன்படுத்துகின்றது. உருவாக்கும் கட்டத்தில் விதி-அடிப்படையில் மாற்றுவது சிக்கலானது என்றும் ஊகிக்க இயலாத விளைவுகளைத் தரும் என்றும் தெரியவந்தபோது எடுத்துக்காட்டு அடிப்படை மேம்படுத்தப்பட்டு ஊடாட்டமாக முன்னேற்றமடைந்தது. உண்மையான பயன்பாட்டின் போது, ஒழுங்குமுறை விதி அடிப்படையை தருவிப்பதற்குமுன் முதலில் பொது எடுத்துக்காட்டு அடிப்படைக்கும் மூல எடுத்துக்காட்டு அடிப்படைக்கும் பொருத்தத்தை முயற்கிக்கும். ஆங்கிலபாரதி-II-இல் தானியக்க முன்

திருத்ததிற்கும் (pre-editiong) பொழிப்புரைக்கும் (paraphrasing) பொதுமைப்படுத்தப்பட்ட மற்றும் கட்டுப்பட்ட பல்சொல் வெளிபடுத்தங்களுக்கும் (generalized and conditional expressions), பெயரிடப்பட்ட பொருள்களை (named entities) அறிந்துகொள்வதற்கும், பொருள்புலத் தனிப்பயனாக்க கருவிகளுக்கும் (domain customization tools), உட்படுத்தப்பட்ட பிழையாய்வுத் தொகுதிக்கும் தானியக்கப் பின் திருத்தத்திற்கான புள்ளியியல்சார் மொழி மாதிரிக்கும் (statistical language model) ஏற்பாடு செய்யப்பட்டிருந்தது.

அனுபாரதி உத்தி அனுபாரதி II-இல் இந்தியை மூலமொழியாகக்கொண்டு பிறமொழிகளுக்கு மொழிபெயப்பதற்கு ஏதுவாக எடுத்துக்காட்டு அடிப்படையின் பொதுமையாக்கம் மூலமொழியைச் சாதிருக்கும் படி பொதுமைப்படுத்தப்பட்டுள்ளது. அனுபாரதி-II-இன் கட்டமைப்பின் மையம் ஒரு பொதுமையாக்கப்பட்ட படிநிலைசார் எடுத்துக்காட்டு அடிப்படையாகும். இந்தியிலிருந்து பிற இந்திய மொழிகளுக்கு எடுத்துக்காட்டு அடிப்படையின் உருவாக்கம் ஒற்றுமையிலாத மொழிகளுடன் ஒப்பிடும் போது மிக எளிமையானதாகும். ஒப்பீட்டளவில் இந்தி பிற எல்லா இந்திய மொழிகளைப் போன்று சுதந்திரமான சொல் வரிசையுள்ள மொழியாகும். உள்ளீடு செய்யப்படும் இந்தி வாக்கியம் சொல்வரிசை வேறுபாடுகளைக் கையளுவதற்காக ஆழமில்லா இலக்கணப் பகுப்பாய்வால் தரப்படுத்தப்பட்ட வடிவாக மாற்றப்படுகின்றது. தரப்படுத்தப்பட்ட இந்தி வாக்கியம் மேல்நிலை தரப்படுத்தப்பட்ட எடுத்துக்காட்டு அடிப்படையுடன் பொருத்தப்படும். பொருத்தம் கண்டுபிடிக்கபடாவிட்டால் ஒரு ஆழமில்லா தொடர்பகுப்பான் (shallow chunker) உள்ளீட்டு வாக்கியத்தை அலகுகளாகத் துண்டுபடுத்தப் பயன்படுத்தப்படும்; இத்துண்டுகள் படிநிலைசார் எடுத்துக்காட்டு-அடிப்படையுடன் பொருத்தப்படும். மொழிபெயர்ப்பு செய்யப்பட்ட தொடர்கள் (chunks) வாக்கிய நிலை எடுத்துக்காட்டு அடிப்படையுடன் பொருத்தத்தால் நிலைபடுத்தப்படும். பிழைஆய்வு தொகுதியும் புள்ளியியல் மொழி மாதிரியும் ஆங்கிலபாரதி II போன்று உட்படுத்தப்பட்டுள்ளன. மனிதப் பின்திருத்தம் முக்கியமாக இந்தியில் இருக்கிற அல்லது மதிப்பிட சிக்கலான அடைகொளி அடைகளைப் புகுத்த வேண்டி செய்யப்படுகின்றது.

4.7.மந்ரா இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை (MANTRA MT System)

மந்த்ரா (MAchiNe assisted TRAnslation tool (MANTRA)) ஆங்கிலப் பனுவலை இந்தியில் மொழிபெயர்க்கும். இது தனிப்பட்ட நிர்வாகம். அரசிதழ் அறிவிப்புகள், அலுவலகக் கட்டளைகள், அலுவலக அறிக்கைகள் மற்றும் சுற்றுக்கள் போன்ற குறிப்பிட்ட பொருண்மைகளங்களில் மொழிபெயர்ப்பு செய்யும். ஆங்கில மற்றும் இந்தி இலக்கணத்தை உருப்படுத்தம் செய்ய மந்ரா சொல்லாக்கப்பட்ட கிளையமைப்பு இணையும் இலக்கணத்தைப் (Lexicalized Tree Ajoining Grammar (LTAG)) பயன்பத்துகின்றது. இது கிளையமைப்பு இணையும் இலக்கணத்தை (Tree Ajoining Grammar (TAG)) பகுப்பாய்வு செய்யவும் ஆக்கம் செய்யவும் பயன்படுத்தும். மந்த்ரா ஒழுங்குமுறைக்கு பனுவல் ஆவணங்கள், பேச்சு அறிதல் நிரல் அல்லது ஒளிஎழுத்துணரி தொகுப்புக்கள் வழியாக உள்ளீடு செய்யலாம்.

தொடக்கத்தில் மந்த்ரா ஒழுங்குமுறை நியமனக்கடிதம், அறிவிப்பு, மைய அரசால் வெளியிடப்படும் சுற்றறிக்கை போன்ற நிர்வாக ஆவணங்களின் மொழிபெயர்ப்புக்காகத் தொடங்கப்பட்டது. பின்னர் இவ்வொழுங்குமுறை நிதி, விவசாயம், உடல்நலப்பராமரிப்பு, தகவல் தொழில்நுட்பம், கல்வி ஆகியப் பொருண்மைக்களத்திற்கும் அரசாங்கப் பொருண்மைக்களத்தில் பொதுவான நோக்கச் செயல்பாட்டிற்கும் நீட்சிசெய்யப்பட்டது. மந்த்ரா தொழில்நுட்பதை அடிப்படையாகக் கொண்ட மந்த்ரா-ராஷ்ட்ரபாஷா என்ற திட்டம் சி-டாக்கால் உருவக்கப்பட்டது. (http://www.cdac.in/html/aai/mantra.asp).

பணித்துறை பொருண்மைக்களத்தில் வாக்கியக் கட்டுமானங்களை ஏற்று பகுத்தாய்வுசெய்து உருவாக்கம் செய்வதற்கு வேண்டி இலக்கணம் சிறப்பாக வடிவமைக்கப்பட்டது. இதுபோன்று அதன் பொருண்மைக் களங்களில் பயன்படுத்தப்படும் ஆங்கிலச் சொற்களின் பொருள்களைக் கையாளுவதற்காக அகராதி பொருத்தமாகக் கட்டுப்படுத்தப்பட்டுள்ளது.

மந்த்ரா-ராஷ்ட்ரபாஷா என்ற ஒழுங்குமுறை இந்தியப் பாராளுமன்றத்தின் மேல்சபையில் ராஜ்யசபா செயலத்திற்காக உருவாக்கப்பட்டது. இது பாராளுமன்ற நடவைக்கைகளை மொழிபெயர்கின்றது. தற்போது ஆங்கிலம்-பெங்காளி, ஆங்கிலம்-தெலுங்கு, ஆங்கிலம்-குஜராத்தி, இந்தி-ஆங்கிலம், இந்தி-பெங்காளி, இந்தி-மராத்தி ஆகிய மொழி இணைகளுக்குத் தொடங்கப்பட்டுள்ளது.

4.8.அனுவாதக் மொழிபெயர்ப்பு ஒழுங்குமுறை

சூப்பர் இன்ஃபோசாஃப்ட் ப்ரைவேட் லிமிடெட்(Super Infosoft Private Limited), தில்லி அனுவாதக் 5.0 என்ற ஆங்கிலம்-இந்தி இயந்திர மொழிபெயர்புக் கருவியை உருவாக்கியது. இது முன் திருத்தத்திற்கு ஆதரவு தரும். இது அதிகாரம், முறைசார்ந்தவை, விவசாயம், மொழியியல், தொழிநுட்பம், நிருவாகம் போன்ற சிறப்பான பொருண்மைகளங்களுக்கு உட்படுத்தப்பட்ட அகராதியைக் கொண்டுள்ளது. ஆங்கிலச்சொல்லுக்கு இந்திப் பொருள் இல்லை என்றால் ஒலிபெயர்ப்பு வசதி தரப்பட்டுள்ளது. மென்பொருள் எந்த விண்டோ குடும்பத்தின் எந்த இயக்கமுறையிலும் (operating system) வேலைசெய்யும்.

அனுவாதக்–ஆங்கிலத்திலிருந்து இந்திய மொழிமொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்யும் ஒரு வலுவான ஒழுங்குமுறை

இது கிளையமைப்பு இணைக்கும் இலக்கணம் (Tree Adjoing Grammar (TAG/டாக்) அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு, புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Statistical Machine Translation (SMT), விதிகள் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (AnalGen/அனல்ஜென்), மற்றும் எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Example Based Machine Translation (EBMT/ஈபிஎம்டி) ஆகியவ இது நான்கு இயந்திர மொழிபெயர்ப்பு தொழில்நுட்பங்களின் ஒருங்கிணைப்பை முன்வைத்துள்ள கூட்டமைப்பு நிறுவனங்களின் கூட்டு முயற்சி ஆகும். ஒரு நிபுணத்துவமான ஆங்கிலத்திலிருந்து இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பு அமைப்பு (English to Indian Language Machine Translation (EILMT)) என்பது ஒரு அதிநவீன தீர்வாகும், இது உரையை ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு மொழிபெயர்க்க அனுமதிக்கிறது.

அனுவாதக்ஷ் இயந்திர மொழிபெயர்ப்பு சி.டி.ஐ.சி./சிடாக் புனே தலைமையிலான 10 இந்திய கல்வி மற்றும் ஆராய்ச்சி நிறுவனங்களின் கூட்டமைப்பால் உருவாக்கப்பட்டது. இந்திய அரசின் மின்னணு மற்றும் தகவல் தொழில்நுட்பத் துறையின் (Departmet of Electronic and Communication (DeitY/டீடி) கீழ் 'இந்திய மொழிகளுக்கான தொழில்நுட்ப மேம்பாடு' (Technological Development of Indian Languges (TDIL/டி.டி.ஐ.எல்) என்ற திட்டத்தின் ஆதரவுடன் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்புகள் 6 மொழி ஜோடிகளுக்கு அதாவது ஆங்கிலத்திலிருந்து இந்தி, பெங்காலி,

மராத்தி, உருது, தமிழ் மற்றும் ஒரியா ஆகிய மொழிகளுக்கு ஒழுங்கமைப்புகள் உருவாக்கப்படுகிறன. ஆரம்பத்தில் சுற்றுலா மறும் உடல்நிலப் பொருண்மைக் களங்கள் எடுத்துக்கொள்ளப்பட்டன. பின்னர் படிப்படையாக பிற பொருண்மைக்களத்திற்கு விரிவுபடுத்தும் நோக்கம் கொண்டுள்ளது.

4.9.உலகளாவிய வலைப்பின்னல் மொழி அடிபடையிலான ஆங்கில இந்தி மொழிபெயர்ப்பு ஒழுங்குமுறை (UNL-based English-Hindi MT system)

பாம்பேயிலுள்ள இந்திய தொழில்நுட்ப நிறுவனம் உலகளாவிய வலைப்பின்னல் மொழியைப் பயன்படுத்தி (Universal Networking Language (UNL) ஆங்கிலம்-இந்தி மொழிபெயர்ப்புக்காக ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்கியது. யுஎன்எல் ஐக்கிய நாடுகளின் பல்கலைக்கழகத்தின் உலகத்திட்டம் ஆகும். இதன் நோக்கம் எல்லா முக்கிய மனித மொழிகளுக்கும் ஒரு இடைமொழியை உருவாக்குவதாகும். பாம்பேயிலுள்ள இந்திய தொழில்நுட்ப நிறுவனம் யுஎன்எல்-இன் இந்திய உறுப்பினராகும். (இந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை பின்வரும் வலைதளத்தில் பயன்படுத்தலாம்: http://www.cfilt.iitb.ac.in/machine-translation/eng-hindi-mt. மேலும் இரண்டு செய்துகாட்டும் ஒழுக்குமுறைகளையும் இந்த வலைதளத்திலிருந்து பயன்படுத்தலாம்: ஒன்று இந்தி-யுஎன்எல் மற்றொன்று யுஎன்எல்-இந்தி மாற்றம். யுஎன்எல் அடிப்படையிலான ஆங்கிலம்-மராத்தி, ஆங்கிலம்-பெங்காளி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் முயற்சிக்கப்பட்டு வருகின்றன.

4.10.மாத்ரா இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை

தேசிய மென்பொருள் தொழிநுட்ப மையத்தின் (National Centre for Software Technology (NCST), மும்பையிலுள்ள (தற்போது சிடாக், மும்பை) அறிவு அடிப்படை கணினி ஒழுங்குமுறைகளின் (Knowledge Based Computer Systems) பிரிவின் இயற்கைமொழி ஆய்வுக் குழு மாத்ராவை உருவாக்கியது. இது ஆங்கிலத்திலிருந்து இந்திக்கு மனித உதவியுடனான மற்றல்-அடிப்படை மொழிபெயர்ப்பு ஒழுங்குமுறையாகும். இது டிடியல்-ஆல் (TDIL) ஆதரிக்கப்பட்ட ஆய்வாகும். மாத்ரா அகராதி மற்றும் அணுகுமுறை பொது

நோக்கத்திற்கானதாகும். இந்த ஒழுங்குமுறை செய்தி, ஆண்டறிக்கைகள், தொழில்நுட்ப தொடர்கள் ஆகிய பொருண்மைகளங்களை மொழிபெயர்க்கப் பயன்படுத்தப்பட்டது.

ஆங்கிலச் செய்திக் கதைகளை இந்தியில் மொழிபெயர்ப்பதற்காக வாக்யா (Vaakya) என்ற கலப்பு அணுகுமுறை ஒழுங்குமுறை என்சிஎஸ்டி (NCST), பாம்பே-இல் உருவாக்கப்பட்டது. இந்த ஒழுங்குமுறை ஒரு தனி வினை வாக்கியத்தைக் கையாள இயலும். மூலமுன்மாதிரி வாக்கியா (Prototype Vaakya) ஒழுங்குமுறை பின்னர் ஆங்கிலச் செய்திக் கதைகளை இந்திக்கு மொழிபெயர்க்கும் சேவையைக் கொடுக்கும்படி மேம்படுத்தப்பட்டு மாற்றியமைக்கப்பட்டுள்ளது.

4.11.ஆங்கில-கன்னடா இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை

ஹைதராபாத் பல்கலைகழகத்தின் கணினி மற்றும் தகவல் அறிவியல் துறை (Computer and Information Sciences) உலகலளாவிய எச்சத்தொடர் அமைப்பு இலக்கணம் (Universal Clause Structure Grammar (UCSG)) வடிவவாதத்தைப் பயன்படுத்தி ஆங்கிலம்-கனடா இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்கியது. இது மாற்றல் அடிப்படையில் அமைந்த அணுகுமுறையாகும். இது அரசு சுற்றுகள் பொருண்மைக்களத்தை மொழிபெயர்க்க பயன்படுத்தப்படும் இந்த ஒழுங்குமுறை கர்நாடக அரசின் நிதியுதவியால் மேற்கொள்ளப்பட்டது. (http://www.languagetechnolies.ac.in/lerc/mat/mat.htm). இது ஆங்கிலம்-தெலுங்கு மொழிபெயர்பு ஒழுங்குமுறையாயகவும் நீட்சி செய்யப்பட்டுள்ளது.

4.12.. இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை (ILTLMT system)

4.12.1. அறிமுகம்

4.12.1.1 நோக்கம்

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்புத் திட்டம் இணையும் அடிப்படையிலான மொழிபெயர்ப்பு திட்டம் 2006 அக்டோபரில் மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் கீழ் தகவல் தொழில் நுட்பத் துறையின் நிதி நல்கையால் நடைபெற்று வருகின்றது. இதன் முதல் நிலை 2006 அக்டோபர் 30-ஆம் நாள் தொடங்கப்பட்டு 2010 ஏப்பிரலில் முடிவுறது. இத்திட்டம் இந்திய மொழியிலிருந்து

மற்றொரு இந்திய மொழிக்கு இயந்திர மொழிபெயர்ப்பை உருவாக்கம் இணையம் அடிப்படையிலான மொழிபெயர்ப்புத் திட்டமாகும். இந்த ஒழுங்குமுறை இரு திசையில் செயல்படும் மற்றும் பொதுவான களத்தில் பின்வரும் மொழிகளுக்கிடையில் செயலாற்றும்.

தமிழ் - இந்தி

தெலுங்கு - இந்தி

மராத்தி - இந்தி

பெங்காளி - இந்தி

தமிழ் - தெலுங்கு

உருது - இந்தி

கன்னடா - இந்தி

பஞ்சாபி - இந்தி

மலையாளம் - தமிழ்

இந்திய மொழி பேசுபவர்கள் இந்த ஒழுங்குமுறையைப் பயன்படுத்துவார்கள். இந்த ஒழுங்குமுறையைக் கற்றல் நுட்பங்களைப் (Machine learning teachiques) பயன்படுத்தி மேலும் இயலும். இந்த ஒழுங்குமுறை இரண்டு குறிப்பிட்ட களங்களுக்கு இயந்திர மொழிபெயர்ப்பை செய்யும்.

பின்வரும் அடிப்படைக் கருவிகள் இந்திய மொழி மொழிப்பெயர்ப்பு ஒழுங்குமுறையின் பகுதியாக உருவாக்கப்பட்டுள்ளது.

1. சொல்வகைப்பாட்டு அடையாளப்படுத்தி (POS Tagger)
2. தொடர் பகுப்பான் (Chunker)
3. உருபனியல் பகுப்பாய்வி (Morph Analyser)
4. வட்டாரச் சொல் குழும (Local word Grouper)
5. மாற்றமைவுப் பகுதி (Transfer Module)
6. இருமொழியப் பொருத்தம் (Billingual Mapping)
7. வாக்கிய நிலை உருவாக்கம் (sentence level generation)
8. வட்டாரச் சொல் குழுமப் பிரிப்பான் (local word group splitter)
9. உருபனியல் உருவாக்கம் (morphological generation)

4.12.1.2 வாய்ப்பு (scope)

இந்த ஒழுங்குமுறை நிலைப்பேறு பெற்ற மொழியில் எழுதப்பட்ட இணைய பக்கங்கள் அல்லது கால இதழ்கள், நாளிதழ்கள் மீது வேலை செய்யும். சரியான தன்மை பயன்படுத்துவோர் திருப்தி (accuracy / user satisfaction) 85 % - 90 % விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிப்பெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்பட்டுள்ளது மற்றும் 80–85 % திறமையுடன் பொதுப் பயன்பாடு இயந்திர மொழிபெயர்ப்பு உருவாக்கப்படுள்ளது.

4.12.1.3. சரியான தன்மை/ பயன்படுதுவோர் திருப்தி (Accuracy/user satisfaction)

85-90 விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிபெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்படும். மற்றும் 80-85 விழுக்காடு திறமையுடன் பொதுப்பயன்பாடு இயந்திர மொழிபெயப்பு உருவாக்கப்படும்.

4.12.1.4 ஒழுங்குமுறைச் சிறப்பு (system specification)

இந்திய மொழிபெயர்ப்புத் திட்டம் இணையத்துடன் தொடர்பு கொண்டு பயன்படுத்துவோருக்குத் தானியக்க மொழிபெயர்ப்பு சேவைஇயைத் தரவேண்டி ஒரு பரிமாறி (Server) இயங்கும். பின்வருவன மென்பொருள் சிறப்பீடுகளாகும்.

Plat form	:	Fedora Core 4 (Kernel – 2.6.11)
Web server	:	Apache – 2.0
Data base	:	GDBM – 1.8
Compiler	:	gcc 4.0

4.12.1.5 நெறிமுறை (methodology)

இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறை பல இயற்கைமொழி ஆய்வு ஆய்வுக் குழுவினர் பங்களிப்பால் உருவாக்கப்பட்டுளது. இந்திய மொழி பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கம் பெரிய செயல்பாடு சிறிய செல்பாடுகளாப் பகுக்கப்பட்டுள்ளது. பங்களிக்கும் ஒவ்வொரு குழுவினரும் ஒன்றோ அதற்கு மேலோ வேளைகளை எடுத்துக்கொண்டு இணையாக வேலை செய்துள்ளனர். பெரும்பாலான பகுதிகள் மொழிச் சுதந்திரமான (language independent engine) இயந்திரமாகவும் மொழிச் சிறப்பான தரவாகவும் (language specific data)

பகுக்கப்பட்டுள்ளது. இது போல் ஒரு மொழிக் குழுவினர் ஒரு குறிப்பிட்ட மொழிக்குத் தரவைத் தருவர். ஒரே இயந்திரம் பல மொழிகளுக்கு வேலை செய்யும்.

4.12.1.5.1 தொகுதிகளாக அமைத்தல்

இந்திய மொழிபெயர்ப்பு ஒழுங்குமுறை பல தொகுதிகளைக் கொண்டது. ஒவ்வொரு பகுதியும் ஒரு தர்க்கம்சார் வேலையைச் செய்யும், பெரும்பாலும் இந்த வேலை சிறியதாகும். இதன் காரணமாக ஏதாவது மாற்றம் ஏற்பட்டால் சிறிய பகுதியில் அதை எளிதாகச் செய்யவியலும்.

4.12.1.5.2 சக்தியின் தரமான அமைப்பு (Shakthi Standard Format (SSF)

எல்லாப் பாகங்களும் (modules)வடிவ அமைப்பு நிர்ணயிக்கப்பட்ட தரவின் ஒழுக்கில் செயல்படும். ஒரு பகுதி அதன் வேலையில் வெற்றி பெற்றால் அது வெளிPயீடு ஒழுங்குப் புதிய சின்னத்தையோ ஆய்வையோ சேர்க்கும். பெரும்பாலான பகுதிகளுக்குத் தனிப்பகுதிகளின் வட்டாரச் சிக்கல்களை உறுதியான கட்டுப்பாட்டில் வைக்க இந்த அணுகுமுறை உதவும்.

4.12.1.5.3 தோல்விகளை நேர்செய்தல்

(SSF) வடிவமைப்பு பகுதியின் ஒவ்வொரு நிலையில் ஏற்படும் தோல்விகளை நேர்செய்ய வசதி செய்யும். எடுத்துக்காட்டாக அதே SSF -இன் மீது அதன் சின்னத்தின் மதிப்பை விளைவிக்கத் தவறினால் கீழ் ஒழுங்குப் பகுதிகள் (down streom modules) தரவின் ஒழுங்கின் மீது தொடர்ந்து செயல்படும்.

4.12.1.5.4 தெளிவு

இந்தியமொழி இயந்திர மொழிப்பெயர்ப்பில் SSF வடிவமைப்பின் பயன்பாடு ஒவ்வொரு பகுதியின் உள்ளீட்டிற்கும் வெளிப்யீட்டிற்கும் முன் நிகழ்ந்திராதத் தெளப்வைப் பெற உதவுகிறது. ஒரு பகுதியின் உரை சார் வெளிப்யீடு மனிதப் பயன்பாட்டிற்கு மட்டுமல்லாமல் அதன் உள்ளீடாக தரவு ஒழுக்கில் அடுத்து வருகிற பகுதியிலும் பயன்படுத்தப்படும்.

4.12.1.5.5 டாஷ்போர்டு

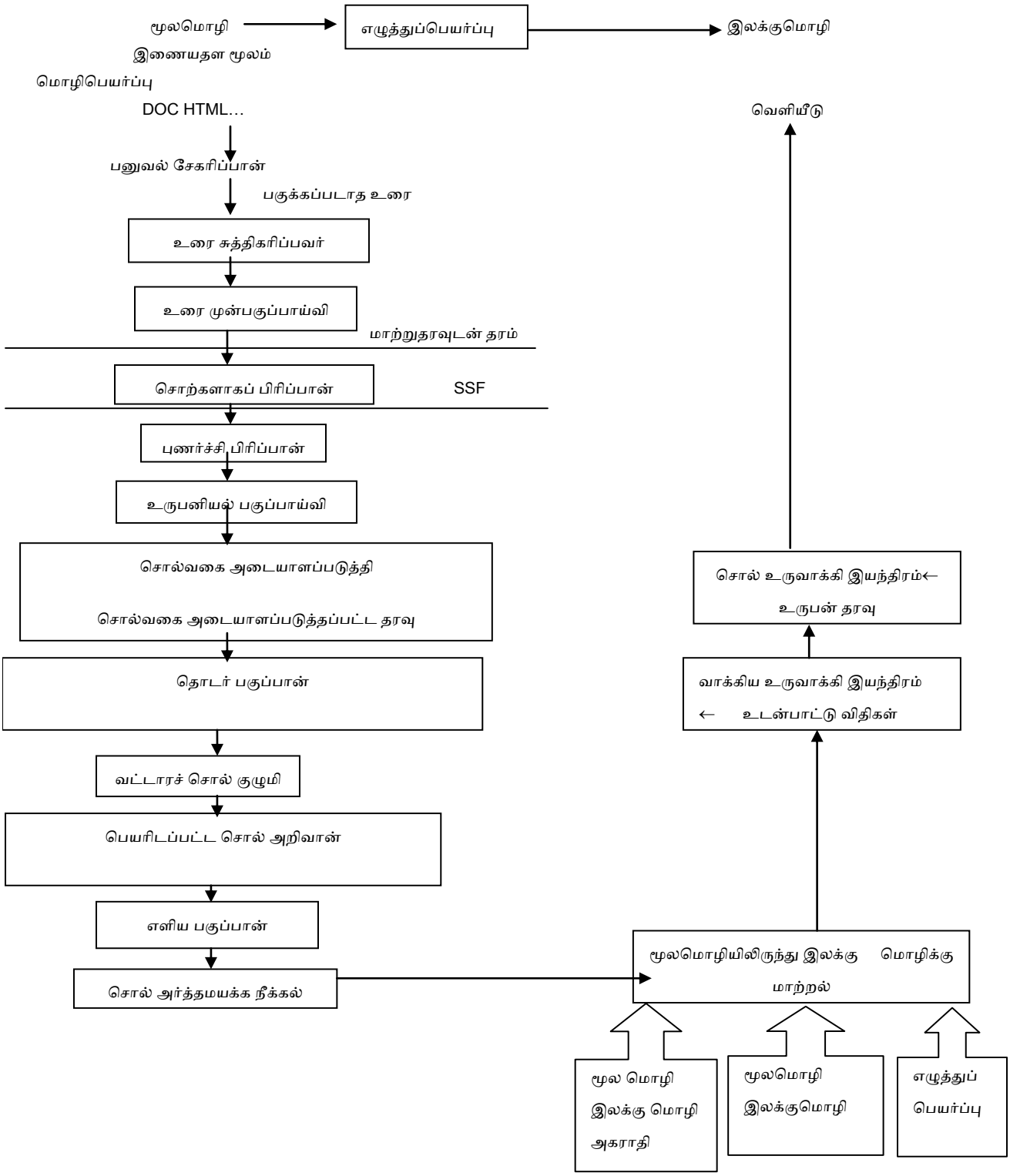
இந்திய மொழி மொழிபெயர்ப்பின் எல்லாப் பகுதிகளும் டாஷ்போர்டின் (dash board) உதவியால் இயங்குகிறது. இது குறிப்பிட்டபடி தரவு ஒழுக்கின் பைப்லைன்களை உருவமைப்பு செய்கிறது. இருப்பினும் பைப் லைன்களுக்குப் பதிலாக எல்லாப் பகுதிகளும் செயல்படும் பங்கிடப்பட்ட நினைவகத்தையோ பிளாக் போர்டையோ (Black Board) நிறுவுவதால்

ஒங்குமுறையை விரைவுப்படுத்த இயலும். இருப்பினும் இது பங்கிடப்பட்ட நினைவாகத்தில் (அல்லது நினைவக உருப்படுத்தத்தில் SSF) இயங்குவதற்கு வேண்டி பகுதிகள் எழுதப்பட்டிருக்கவேண்டும் என்று அவசியப்படும்.

4.12.1.6. முழுவிளக்கம்

இந்திய மொழியிலிருந்து இந்திய மொழிக்கான மொழிபெயர்ப்பு ஒழுங்குமுறை ஆய்தல் - மாற்றியமைத்தல் - உருவாக்குதல் என்ற மேற்கோள் வாய்ப்பாடு அடிப்படையில் செயல்படும். முதலாவது மூலமொழியின் ஆய்வு செய்யப்படும். பின்னர் இலக்குமொழிக்குச் சொற்றொகையும் ஆயப்பட்ட அமைப்பும் மாற்றம் செய்யப்படும். இறுதியாக இலக்கு மொழி உருவாக்கப்படும்.

இந்திய மொழிகள் ஒற்றுமையுள்ளதாக இருப்பதாலும் இலக்கண அமைப்புகளைப் பங்கீட்டுக் கொள்வதாலும் ஆழமற்ற பகுத்துக் குறித்தல் செய்யப்படும். மாற்றமைவு இலக்கணப் பகுதி எளிமையாக வைக்கப்படும். பொருண்மை களச் சிறப்புப் பெயரிடப்பட்ட சொற்களின் புரிந்து கொள்வான்களாலும் பொருத்தமான அகராதி போன்றவற்றாலும் கையாளப்படும். ஆழமற்ற பகுத்துக் குறித்தலானது, உருபனியல் ஆய்வு, சொல் வகைப்பாடு அடையாளப்படுத்தல் மற்றும் தொடர்பகுத்தல் என்பனவற்றை உள்ளடக்கும். முதலாவது (உருபனியல் பகுத்தாய்வு) விதி அடிப்படையிலானது, இரண்டாவது (சொல் வகைப்பாடு அடையாளப்படுத்தல்) புள்ளியியல் அடிப்படையிலானது, மூன்றாவது (தொடர்ப்பு) விதி அடிப்படை மற்றும் புள்ளியியல் அடிப்படை இரண்டும் இணைந்தது. முழு ஆய்வு வேலையும் பல பாகங்களைப் பிரிக்கப்பட்டுள்ளது. ஒவ்வொரு பாகமும் சிறிய தர்க்க வேலையைச் செய்யும். இந்த ஒழுங்குமுறையின் முழு அமைப்பும் கீழே தரப்பட்டிருக்கும் படத்தில் காட்டப்பட்டுள்ளது.


மூலமொழி
இணையதள மூலம்
மொழிபெயர்ப்பு
DOC HTML…
எழுத்துப்பெயர்ப்பு
இலக்குமொழி
வெளியீடு
பனுவல் சேகரிப்பான்
பகுக்கப்படாத உரை
உரை சுத்திகரிப்பவர்
உரை முன்பகுப்பாய்வி
மாற்றுதரவுடன் தரம்
சொற்களாகப் பிரிப்பான்
SSF
புணர்ச்சி பிரிப்பான்
உருபனியல் பகுப்பாய்வி
சொல்வகை அடையாளப்படுத்தி
சொல்வகை அடையாளப்படுத்தப்பட்ட தரவு
தொடர் பகுப்பான்
வட்டாரச் சொல் குழுமி
பெயரிடப்பட்ட சொல் அறிவான்
எளிய பகுப்பான்
சொல் அர்த்தமயக்க நீக்கல்
சொல் உருவாக்கி இயந்திரம்←
உருபன் தரவு
வாக்கிய உருவாக்கி இயந்திரம்
← உடன்பாட்டு விதிகள்
மூலமொழியிலிருந்து இலக்கு மொழிக்கு
மாற்றல்
மூல மொழி
இலக்கு மொழி
அகராதி
மூலமொழி
இலக்குமொழி
எழுத்துப்
பெயர்ப்பு

4.12.2.ஒழுங்குமுறையின் அமைப்பு

இந்திய மொழியிலிருந்து இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை பகுப்பாய்வு-மாற்றல்-உருவாக்கம் என்பதன் அடிப்படையில் இருக்கும். முதலில், மூல மொழியின் பகுப்பாய்வு செய்யப்படும்; பின்னர் இலக்கு மொழிக்கான சொற்றொகை மற்றும் பகுப்பாய்வு கட்டமைப்பு ஒரு மாற்றல் மேற்கொள்ளப்படும்; இறுதியாக இலக்கு மொழி உருவாக்கப்படும்.

இந்திய மொழிகள் ஒத்தவை மற்றும் இலக்கண கட்டமைப்புகளைப் பகிர்ந்து கொள்வதால், மேலோட்டமான பாகுபடுத்தல் மட்டுமே செய்யப்படுகிறது. மாற்றல் இலக்கண கூறு எளிமையாக வைக்கப்படும். பொருண்மைக்கள குறிப்பிட்ட அம்சங்கள் பொருத்தமான பெயர்-இருப்பான் புரிவான்கள் பொருத்தமான அகராதி உள்ளீடுகள் போன்றவற்றால் கையாளப்படும்.

மேலோட்டமான பாகுபடுத்தல் (ஆழமில்லாப் பகுப்பாய்வு (shallow parsing)) மூன்று முக்கிய பணிகளை உள்ளடக்கும்: உருவவியல் பகுப்பாய்வு, சொல்வகைப்பாட்டு அடையாளப்படுத்தல் மற்றும் தொடர் உறுப்பு பகுத்தல். முதலாவது விதி அடிப்படையிலானதாக இருக்கும், இரண்டாவது புள்ளிவிவரம் மற்றும் மூன்றாவது இரண்டின் கலவையாகும். ஒட்டுமொத்தமாக செயலாக்க பணி பல தொகுதிகளாக பிரிக்கப்பட்டுள்ளது, அவை ஒவ்வொன்றும் பொதுவாக ஒரு சிறிய தருக்க பணி. அமைப்பின் ஒட்டுமொத்த கட்டமைப்பு படம் முன்னர் கொடுக்கப்பட்டுள்ளது. முக்கியமான தொகுதிகள் ஒவ்வொன்றும் கீழே விவரிக்கப்பட்டுள்ளன. ஒவ்வொரு தொகுதி எதிராக குறிப்பிடப்பட்ட முக்கிய பணியை செய்கிறது. இருப்பினும், முக்கிய பணியைத் தவிர, இது சில முன் செயலாக்கம் மற்றும் பின் செயலாக்கத்தையும் மேற்கொள்ளக்கூடும். எடுத்துக்காட்டாக, தொடர் பகுப்பான் தொகுதி (chunker module) தொடர்களை அடையாளம் காட்டுகிறது. இதனால், தொடர்பகுப்பான் தொகுதி தொடர்பகுக்கும் (chunking) முக்கிய பணியை செய்கிறது. இருப்பினும், இது போஸ்ட்ரோசெசிங்கையும் செய்யக்கூடும்; இதில் தொடர் அடையாளம் காணப்படுகிறது மற்றும் தலையின் பண்புக்கூறுகள் தொடரின் பண்புக்கூறுகளாக நகலெடுக்கப்படுகின்றன.

ஒவ்வொரு பகுதியும் அதற்கு நேரே தரப்பட்டுள்ள முக்கிய வேலையைச் செய்கின்றன. இருப்பினும் முக்கிய வேலை தவிர அவை சில முன்னாய்வையும் (Pre-processing) பின்னாய்வையும் (Post-processing) செய்கின்றன. எடுத்துக்காட்டாகத் தொடர்பகுப்பான் பகுது தொடர் பகுதிகளைக் கண்டுபிடிக்கின்றன. இவ்வாறு தொடர்பகுப்பான் பகுது முக்கிய

வேலையான தொடர் பகுத்தலைச் செய்கின்றது. இருப்பினும் தொடர் பகுதியின் தலைமையைக் கண்டுபிடிக்கும் பின்னாய்வைச் செய்கின்றது. மேலும் தலைமைப் பண்புக் கூறுகளைத் தொடர்பகுதியின் பண்புக்கூறுகளாக நகல் செய்கின்றது.

4.12.3. செயல்முறை விளக்கம்

ILMTஇன் அனைத்து தொகுதிகளும் பொதுவான தரவு உருப்படுத்தத்தில் செயல்படுகின்றன, அதாவது SSF (Sakthi Standard Format). ஒரு வாக்கியத்திற்கான SSF உருப்படுத்தம் கிளையமைப்புகளின் வரிசையைக் கொண்டுள்ளது. ஒவ்வொரு கிளையமைப்பும் ஒன்று அல்லது அதற்கு மேற்பட்ட தொடர்புடைய கணுக்களால் ஆனது. ஒரு கணுவிற்க்கு ஒழுங்குமுறைப் பண்புகள் உள்ளன; அவை சரியான பெயரால் (propname) மற்றும் சரியான மதிப்பால் (propvalue) வழங்கப்படுகின்றன. அனைத்து தொகுதிகளும் உள்ளீட்டு தரவைப் SSF வடிவத்தில் படித்து வெளியீட்டுத் தரவை உருவாக்குகின்றன. ஒரு தொகுதி அதன் செயல்பாட்டில் வெற்றிபெற்றால் அது பொதுவாக கணுக்களின் ஒழுங்குமுறைப் பண்புகளை அல்லது பண்புக்கூறு அமைப்புகளின் பயனர் வரையறுக்கப்பட்ட அடைகளை மாற்றியமைக்கும் (நான்காவது ஒழுங்குமுறை செயல்திறன்). இவ்வாறு குழாய்வழியில் கட்டமைப்பு வரைபடத்தில் காட்டப்பட்டுள்ள தொகுதிகள், ஒவ்வொரு தொகுதியும் குறிப்பிட்டதை சேர்க்கிறது அல்லது மாற்றியமைக்கிறது. முந்தைய தொகுதிகளால் உருவாக்கப்பட்ட பிற அடைகள் பண்புகளை. பிற பண்புக்கூறுகள் குழாய்த்திட்டத்தில் முன்னோக்கி கொண்டு செல்லப்படுகின்றன. செயல்முறை விளக்கங்கள் ஒவ்வொரு தொகுதிக்கூறுகளும் கீழே விவரிக்கப்பட்டுள்ளன.

4.12.4. தனிப்பட்ட தொகுதிகளின் விவரக்குறிப்பு

ஒவ்வொரு தொகுதிகளின் விவரக்குறிப்பும் கீழே கொடுக்கப்பட்டுள்ளது. ஒவ்வொரு தொகுதியின் விவரக்குறிப்பும் அது எந்த பண்பு-மதிப்பைக்/மதிப்புகளை உள்ளீடாகப் பயன்படுத்துகின்றது மற்றும் எந்த பண்பை வெளீடாகப் பயன்படுத்துகின்றது என்பதைக் குறிக்கிறது.

4.12.4.1. முன்செயலாக்கி

முன்செயலாக்கி (preprocessor) ஐ.எல்.எம்.டி ஒழுங்குமுறைக்கான ஒரு இடைமுகத்தை வலைக்கு (web) வழங்கும். இந்த தொகுதி வலைச் சேவையகத்திலிருந்து (வெப்சர்வர் (webserver)) உள்ளீட்டு உரையை சேகரிக்கும்; இது பயனரிடமிருந்து வரும் HTTP கோரிக்கை வடிவத்தில்

இருக்கும். இது உரையைச் சுத்தம் செய்யும். சுத்தம் செய்வது அதன் உயர்வான வடிவமைப்பின் வலை உள்ளடக்கத்தை அகற்றுவதை உள்ளடக்கும். இது விருப்பமாக ஒரு எழுத்து குறியமாக்கத்தை மற்றொன்றாக மாற்றக்கூடும். குறைந்தபட்சம் இது யூனிகோட், டபிள்யூ.எக்ஸ்.ரோமன், மற்றும் ISCII என்பனவற்றை ஆதரிக்கும்.

4.12.4.2. டோக்கனாக்கி (Tokenizer)

டோக்கனாகி கொடுக்கப்பட்ட உள்ளீட்டு உரையை உள்ளீட்டு உரையின் ஒவ்வொரு வாக்கியத்தின் முடிவிலும் ஒரு வாக்கியக் குறிப்பானுடன் டோக்கன்களின் வரிசையாக மாற்றுகிறது (சொற்கள், நிறுத்தற்குறிகள் மற்றும் பிற குறியீடுகள்). வெளியீடு SSF வடிவவில் உற்பத்திசெய்யப்படுகிறது. டோக்கனாக்கியின் உள்ளீடு ஒரு சிஎம்எல் (CML) உரையாகவும் இருக்கலாம். வாக்கியக் குறி உரை உள்ளீட்டு தரவுக்கு பயன்படுத்தப்படும் எழுத்து உருப்படுத்த வகையைப் பொறுத்தது. எழுத்துக் குறியாக்கதைப் பொறுத்து இது இவாற்றில் ஒன்றாக இருக்கலாம் (முழு நிறுத்தம், ஆச்சரியத்தின் அடையாளம், விசாரணைக்கான அடையாளம் மற்றும் பூர்ணா விராம் ஆகியவவை). ஒவ்வொரு மொழிக்கும் சிறப்பு டோக்கன்களைக் கையாள டோக்கனாக்கி கட்டமைக்கப்படும்; எடுத்துக்காட்டாக தமிழ் மொழியில் ஸ்ஷ் போன்ற டோக்கன்கள் பிரிக்கப்படாது. உரைக்குள் இருக்கும் ரோமான் குறியீடுகள் அப்படியே இருக்கும், மேலும் வெளியீட்டில் “@” அடையாளத்தால் முன்னதாக இருக்கும்.

4.12.4.3. சந்தி/புணர்ச்சி பிரிப்பான் (Sandhi splitter)

இந்த தொகுதி சந்தி நடைபெறும் மொழிகளுக்கு பயன்படுத்தப்படும். இது ஒரு கூட்டு வார்த்தையை உள்ளீடாக எடுக்கும், அதன் வெளியீடு எளிய சொற்களாக இருக்கும். பல இந்திய மொழிகளுக்கு சந்தி ஏற்படுகிறது மற்றும் இரண்டு அல்லது அதற்கு மேற்பட்ட சொற்கள் ஒரு கலவையான சொல் ஒன்றிணைக்கப்படுகின்றன (சந்தி). அத்தகைய சொற்களில் உருவவியல் பகுப்பாய்வு செய்ய, முதலில் கலப்பு வார்த்தையை இரண்டு அல்லது அதற்கு மேற்பட்ட எளிய சொற்களாகப் பிரிக்க வேண்டியது அவசியம். சில சொற்களுக்கு சந்தி பிரிப்புக்குப் பின்னரும் உருபனியல் பகுப்பாய்வு முடிந்த பிறகும் சாத்தியமில்லை. அத்தகைய வார்த்தைகளுக்கு சந்தி பிரிப்பான் எஃப்எஸ்ஸில் (FS) உள்ள சொல் வகை பண்புக்கூறு அறியப்படாத (“unk”) என அமைக்கப்பட்டிருப்பதால் இந்த வார்த்தையை விட்டு விடுகிறது.

4.12.4.4. உருபனியல் பகுப்பாய்வி (Morph analyzer)

ஒரு சொல் தரப்படும் போது உருபனியல் ஆய்வி அதன் வேரையும் சொல்லின் இலக்கணப் பண்புக்கூறுகளையும் அடையாளங்கண்டு கொள்கிறது. திரிபு வளம் இல்லாத மொழிகளுக்கு எல்லாச் சொல் வடிவுகளையும் கொண்ட எளிய அகராதி நோக்கீடு போதுமானது. தமிழ், தெலுங்கு போன்ற திரிபு வளமுள்ள மொழிகளுக்கு அகராதியை உருவாக்குவது எளிதல்ல. இது பெரிய சேகரிப்பையும் உயர்ந்த செயல்பாட்டுக் கணிப்பையும் வேண்டும். இதற்கு நல்ல மாற்று வழி வேர்ச்சொற்களின் அகராதியை வைத்துக் கொண்டு உருபனியல் ஆய்வியைப் பயன்படுத்தி சொல் வடிவை வேராகவும் இலக்கணப் பின்னுருபுகளாகவும் பிரிப்பதாகும். மேற்கோள்/முன்னுதாரண வாய்ப்பாடு அடிப்படையிலான அணுகுமுறைகள் (Paradigm based approaches) இந்திய மொழிகளுக்கு நல்ல வெற்றி மதிப்பீட்டைத் தருகிறது.

4.12.4.5. சொல்வகைப்பாடு அடையாளப்படுத்தி

சொல்வகை அடையாளப்படுத்தல் வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் ஒரு சொல் வகைப்பாட்டைத் தருவதாகும். வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் பெயர்கள், வினைகள், பெயரடைகள், வினையடைகள் என்று அடையாளம் தருவது ஒரு வாக்கியத்தில் ஒவ்வொரு உறுப்பு பின் ஒரு பங்களிப்பை ஆய்வதற்கு உதவும். சொல்வகைப்பாட்டு அடையாளப் படுத்துவதற்கு விதி அடிப்படை, புள்ளியியல் அடிப்படை போன்ற பல அணுகுமுறைகள் இருக்கின்றது. இங்கு புள்ளியல் அணுகுமுறை பின்பற்றப்படுகின்றது (HMM அடிப்படை).

4.12.4.6. தொடர்கூறு பகுப்பான்

இந்த தொகுதி நான்கு பாகங்களாகப் பிரிக்கப்பட்டுள்ளது.

4.12.4.6.1. தொடர்கூறு பகுத்தல்

தொடர்கூறு பகுத்தல் (Chunking) ஒரு வாக்கியத்தில் உள்ள எளிய பெயர்த்தொடர்கள், வினைக்குழுமங்கள், பெயரடைத் தொடர் மற்றும் வினையடைத் தொடர்களை அடையாளம் காண்பது ஆகும். இது தொடர்கூறு களின் எல்லை மற்றும் புலக்குறிப்பை அடையாளங்காண்பதை உட்படுத்தும்.

4.12.4.6.2. சீரமைப்பு

சீரமைப்பு (pruning) CAT_, யூகித்தல் போன்றவற்றுடன் இணக்கத்தன்மையை சரிபார்ப்பதன் மூலம் உருனியல் பகுப்பாய்வுக்குப் பிறகு கிடைக்கும் மிகவும் பொருத்தமான பண்பு

கட்டமைப்புகளை அடையாளம் காண்கிறது. இந்தத் தொகுதிகள் மேலும் மூன்றாக பிரிக்கப்படுகின்றன.

4.12.4.6.2.1. உருபு சீரமைப்பு

இது lcat (சொல்சார் வகைப்பாடு (lexical category)) மதிப்பு CAT_ மதிப்புடன் பொருந்தக்கூடிய பண்புக்கூறு கட்டமைப்பை எடுக்கும். அந்த பண்புக்கூறுகள் அனைத்தும் கட்டமைப்புகள் கொடுக்கப்பட்ட டோக்கனுக்கான சாத்தியமான வெளியீடுகளாக அவை தக்கவைக்கப்படுகின்றன. மீதமுள்ள பண்புக்கூறு கட்டமைப்புகள் இந்த தொகுதி மூலம் சீரமைக்கப்படும் (அகற்றப்படும்). lcat ஆனது CAT_ உடன் பொருந்துகிற எந்த பண்புக்கூறு கட்டமைப்புகளும் இல்லை என்றால், பின்னர் அனைத்து பண்புக்கூறு கட்டமைப்புகளும் தக்கவைக்கப்படுகின்றன மற்றும் புதிய பண்பு மதிப்பு இணை poslcat = ”NM” ஒவ்வொரு பண்புக்கூறு கட்டமைப்பிலும் சேர்க்கப்பட்டுள்ளது. NM என்பது “பொருந்தவில்லை" என்பதைக் குறிக்கும்.

4.12.4.6.2.2. உருபை யூகிக்கவும் (Guess Morph)

இந்தத் தொகுதி, உருபு சீரமைப்பால் எஞ்சியிருக்கும் பண்புக்கூறு கட்டமைப்புகளின் பன்மடங்கு எண்ணிக்கையைக் குறைக்க பட்டறிவைப் பயன்படுத்துகிறது. இந்த தொகுதி ஒவ்வொரு தனி மொழி குழுக்களால் அந்தந்தத் மொழிகளுக்காக வரையறுக்கப்படும்.

4.12.4.6.2.3.ஒரு உருபைத் தேர்ந்தெடுக்கவு (Pick one morph)

இது கொடுக்கப்பட்ட தேர்வு வரையறையின் அடிப்படையில் ஒரே ஒரு பண்புக்கூறுகட்டமைப்பை மட்டுமே எடுக்கும். அது இயல்பாகவே நடக்கும் முதல் பண்புக்கூறு கட்டமைப்பைத் தேர்ந்தெடுக்கும்.

4.12.4.6.3. தலை கணக்கீடு (Head Computation)

இந்தத் தொகுதி தொடரின் தலையை கணக்கிடுகிறது. ஒரு குழந்தைக் கணு தொடரின் தலையாக அடையாளம் காணப்படுகிறது. *பெயர்* (*name*) என அழைக்கப்படும் ஒரு புதிய பண்புக்கூறு பண்புடன் *பெயர்* ஆக இந்த குழந்தைக் கணுவுடன் சேர்க்கப்பட்டுள்ளது. பண்பு *பெயர்* பின்னர் தன்னிச்சையான ஆனால் தனித்துவமான வாசிப்புக்காக வேருக்கு அருகில் ஒலிக்கும் கோர்வைக்கு பொருத்தப்படுகிறது. தொடரின் கணு, தலை குழந்தையின் பண்புக்கூறுகளைப் உரிமையாகப் பெறுகிறது. *‘பெயர்’* தவிர அனைத்துப் பண்புக்கூறுகளும்

தலை-குழந்தையிலிருந்து பெற்றோர் தொடருக்கு நகலெடுக்கப்படுகின்றன. *தலை* என்ற ஒரு புதிய பண்பு தொடர் கணுவின் பண்புக்கூறுக்குச் சேர்க்கப்பட்டுள்ளது. அதன் மதிப்பு பெயர் தலை-குழந்தைக்கு ஒதுக்கப்பட்ட பெயர்-கோர்வை ஆகும்.

4.12.4.6.4. தலைப் பண்புக்கூறுகளை மரபுரிமையாகப் பெறு (Inherit head features)

'பெயர்/name' தவிர் எல்லா பண்புக்கூறுகளு தலை-குழந்தையிலிருந்து (head-child) பெற்றோர் தொடருக்கு (parent chunk) நகலெடுக்கப்படுகிறன.

4.12.4.7. குறிப்பிட்ட இடம்சார் சொல் குழுமி/பிரிப்பான் (Local Word Grouper)/Splitter

4.12.4.7. 1.குறிப்பிட்ட இடம்சார் சொல் குழுமி (Local Word Grouper)/விபக்தி கணினியாக்கம்

இடம்சார் சொல் குழுமி விபக்தி கணினியாக்கத்தின் (Vibhakti Computation) தொழில்நுட்ப வேலையைச் செய்யும் தனிச்சொற்களாக வரும் துணை வினைகள் தனி பொருளடக்கச் சொற்களாகக் கருதப்படுவதில்லை. ஏனென்றால் அவை முதன்மை வினையின் தலைமையாக இருக்கும் தொடர்பகுதியின் இலக்கணப் பண்புக் கூறுகளைச் சிறப்பீடு செய்யும். இது இந்தியில் தனிச் சொற்களாக வருகின்ற பெயர்விபக்தி என்று அழைக்கப்பெறிகின்ற பின்னுருபுகளுக்கும் உண்மையாகும். இடம்சார் சொல் குழுமி பொருத்தமான கலவை நிலை மாறும் கூட்டுநிலை சொல் குழும ஆக்க விதிகளைத் திறமையாகக் கையாளும். முக்கியமான வேலை இடம்சார் தகவல் அடிப்படையில் செயல்பாட்டுச் சொற்களைப் (Functional Words) பொருளடக்கச் சொற்களுடன் (Content Words) குழுமுவதாகும். இந்த வேலையை நிறைவேற்ற பயன்படுத்தப்படும் ஒரு செயல்முறை கிரியா ரூப்ப அட்டவணைகளைப் (Kriya rupa Charts) பயன்படுத்துவதாகும். இந்த அட்டவணைகள் ஒரு தனி செயல்பாட்டைக் குறிக்கும் தொடர்ச்சியான வினைகளிலிருந்து உருவாக்க வேண்டிய குழுமங்களைச் சிறப்பீடு செய்யும்.

இந்தத் தொகுதி பெயர்/வினை (noun/verb) தொடர்களின் (chunks) வேற்றுமை/கால-வினையாற்றுவகை-வினைநோக்கு (case/tam (tense-aspect-mode)) பண்புக்கூறுகளை கணினியாக்கம் செய்து அவற்றை FS உடன் சேர்க்கின்றது.

4.12.4.8. பெயரிடப்பட்ட இருப்புபொருளை அறிதல் (Named Entity Recognizer (**NER))**

மொழிப் பகுப்பாய்வு கட்டங்களில் பெயரிடப்பட்ட இருப்பானை அறிவான் (Named Entity Recognizer) முக்கிய பங்கு வகிக்கிறது. பெயரிடப்பட்ட இருப்பானை அறியும் செயல்பாடு ஆவணத்தில் குறிப்பிடப்பட்டுள்ள இருப்பான்கள் கண்டறியப்படுவதை வேண்டும்; அவற்றின்

அர்த்தமயக்கம் நீக்கப்படவேண்டும்; இருப்பானுக்கு ஒதுக்கப்பட்டுள்ள பண்புகள் தெரிந்தெடுக்கப்படவும் ஒரு அடையாளத்தால் உருப்படுத்தம் செய்யப்படவும் வேண்டும். ஒரு இருப்பான், உலகில் உள்ள ஒரு பொருள் அல்லது பொருட்களின் தொகுப்பாகும். பெயரிடப்பட்ட இருப்பானில் படிநிலை அமைப்பு மூன்று முக்கிய வகுப்புகளாக பிரிக்கப்பட்டுள்ளது: பெயர் (name), நேரம் (time) மற்றும் எண்சார் (number) வெளிப்பாடுகள். NER இயந்திரத்தின் செயல்பாடு ஒரு இருப்பான் எந்த வகுப்பில் வருகின்றது என்பதைக் கண்டறிந்து தொடர்புடைய அடையாளத்தைக் கொடுப்பதாகும்.

4.12.4.9. எளிய பகுப்பான் (Simple Parser)

இந்தத் தொகுதி ஒரு வாக்கியத்தில் சில பகுதிகளுக்கு இடையிலான உறவை அடையாளம் காட்டுகிறது. இந்த உறவுகளில் சில பகுப்பாய்வு செய்யப்படும் மொழிசார்ந்து இருக்கும்.

4.12.4.10. சொற்பொருண்மை மயக்கநீக்கம் (Lexical Sense Disambiguation)

உருபனியல் பகுப்பாய்வியால் அடையாளம் காணப்பட்ட வேர்ச் சொற்கள் பொருண்மை மயக்கநீக்கம் செய்யப்படவேண்டும் மற்றும் உள்ளீட்டு உரையில் உள்ள ஒவ்வொரு சொல்லின் அர்த்தத்தின் அடையாளம் கண்டுபிடிக்கப்படவேண்டும். இந்தச் சொற்கள் பன்மொழி அகராதியில் அவற்றின் இலக்கு மொழி நிகரன்களால் இடமாற்றம் செய்யப்படுகின்றன.

4.11.4.11. மூல மொழியிலிருந்து இலக்குமொழிக்கு மாற்றல்

மூல மொழியிலிருந்து இலக்கு மொழிக்கு மாற்றுவதைச் செய்வதில் மூன்று தொகுதிகள் உள்ளன:

a) மொழி இணைகளுக்கு குறிப்பிட்ட மொழி சுதந்திர மாற்றல் இயந்திரம் மற்றும் மாற்றல் இலக்கணத்தை உள்ளடக்கிய மாற்றல் கட்டமைப்புத் தொகுதி (Transfer structure module).

b) பொருத்தமான இருமொழிய ஒருபொருள்பன்மொழிகள் அல்லது இருமொழி சொல் அகராதிகள் தேடும் சொல் பதிலீட்டைச் செய்யும் சொல்சார் பதிலீட்டுத் தொகுதி (Lexical sustitution module).

c) மொழிபெயர்க்கப்படாத சொற்களுக்கு எழுத்து நிலை ஒலிபெயர்ப்பைச் செய்யும் ஒலிபெயர்ப்பு தொகுதி (Transliteration module).

4.12.4.11.1. மாற்றல் எந்திரத் தொகுதி (Transfer Engine Module)

மொழியாய்வில் வாக்கியங்கள் அவற்றின் தொடரியல் அமைப்பை அடையாளம் காண்பதற்காகப் பகுக்கப் குறிக்கப்பெறும். இந்திய மொழிகளுக்குள் வேறுபாடுகளை விட

ஒற்றுமைகள்தான் அதிகம். எடுத்துக்காட்டாகத் தமிழ், இந்தி மொழியினை முழு பகுத்தலை வேண்டாது. இத்திட்டத்தில் இயந்திர மொழிபெயர்ப்பு முழு வாக்கியப் பகுப்பான் இல்லாது நிறைவேற்றப்படும். மூலமொழியின் அமைப்பிற்கு இலக்குமொழி அமைப்பில் நிகரான அமைப்பு இல்லாவிட்டால் அமைப்பு சார் மாற்றம் தேவைப்படும். ஒரு பகுதி பகுத்தல் (partial parse) அல்லது ஆழமற்ற பகுத்தல் (shallow parse) மாற்றத்திற்கு உட்படுத்தப்படும் வாக்கியங்களின் குறிப்பிட்ட உறுப்புகளை அடையாளம் கண்டால் போதுமானது.

மொழிகளுக்கிடையில் உள்ள தொடரியல் வேறுபாடு கலவை வாக்கிய கட்டுமானங்களில், குறிப்பாக எச்சத்தொடர்களில் காணப்படும். இது மாற்றல் இலக்கணத்தால் (transfer grammar) இணைக்கப்படும். இந்த உட்கூறு மூன்று துணைத் தொகுதியைக் கொண்டிருக்கும்: மாற்றல் இலக்கணம் (transfer grammar), சொல்சார் மாற்றல் (Lexical transfer), எழுத்துப்பெயர்ப்பு (transliteration). எழுத்துப்பெயர்புத் தொகுதி இந்திய மொழிகளுக்குள் (உருது மொழியை உள்ளடக்கி) எழுத்துப்பெயர்ப்பை நிறைவேற்ற இயலும் ஒரு தொகுதி உருவாக்கப்பட வேண்டும். எழுத்துப்பெயர்ப்பு சொல்லையோ சொற்களையோ படிப்பவரின் எழுத்துருவுக்கு மாற்றுகிறது. எடுத்துக்காட்டாக ஒருவருக்கு இந்தி தெரிந்திருந்தால் தேவநாகரியில் வங்காள உரையைப் படிக்க இயலும் என்றால் அதன் பொருளின் சில பாகங்களை அவரால் புரிந்து கொள்ள இயலும். மொழிபெயர்ப்பு ஒரு சொல்லுக்கோ ஒரு தொடர் கூறுக்கோ தோற்றுப்போனால் எழுத்துப்பெயர்ப்பு படிப்பவரை படிக்கவும் புரிந்து கொள்ளவும் அனுமதிக்கும். இந்திய மொழிகள் பல எண்ணிக்கையிலான சொற்களைப் பங்கிட்டுக் கொண்டுள்ளன. எனவே படிப்பவருக்கு ஏற்றவாறு எழுத்துருமாற்றம் செய்தால் அவரால் மூலமொழியில் உள்ள உரையிலிருந்து சிலவற்றைப் புரிந்து கொள்ள இயலும்.

4.12.4.11.2. சொல்சார் மாற்றல் எந்திரம் (Lexical Transfer Engine)

உருபனியல் பகுப்பாய்வியால் அடையாளம் காணப்பட்ட வேர்ச் சொற்கள் இலக்கு மொழிக்கான நிகரனுக்காக ஒருபொருள்பன்மொழிய அகராதியில் பார்க்கப்படுகின்றன. இந்த அகராதி இலக்கு மொழியின் வேர்சொல் மற்றும் அதன் சொல்வகைப்பாடு (பெயர்ச்சொல், வினை போன்றவற்றை) மற்றும் பிற தேவையான தகவல்களைக் கொண்டுள்ளது. இந்த நிலை மூல மொழி இலக்கணப் பின்னொட்டுகளுக்கான இலக்கு மொழி நிகரன்களை அடையாளம் காண அகராதிகளைப் பயன்படுத்துகிறது.

4.12.4.11.3. எழுத்துப்பெயர்ப்பு

எழுத்துப் பெயர்ப்புத் தொகுதி மூலமொழி யூனிகோடில் இருந்து "@" சின்னத்துடன் தொடங்கும் வார்த்தையை இலக்குமொழியை யூனிகோட் வடிவத்தில் மாற்றுவதற்கு எழுத்துப்பெயர்ப்பு தொகுதி பயன்படுத்தப்படுகிறது.

4.12.4.12. இலக்கு மொழிப் பண்புக்கூறுகளைஇடு

இந்த தொகுதியின் முதன்மைப் பணி, பெயர்ச்சொற்களின் பாலின பண்புக்கூறை இலக்கு மொழிக்கு ஏற்ப செருகுவதாகும். "paMKa" போன்ற சில பெயர்ச்சொற்கள் தெலுங்கில் பெண்பால், இந்தியில் ஆண்பால். எனவே, இதற்காக இலக்கு மொழியில் (இந்தி) பெண்பால் பெயர்ச்சொற்களின் பட்டியல் சேகரிக்கப்பட்டது. இந்த தொகுதி பட்டியலில் உள்ள அனைத்துப் பெயர்ச் சொற்களையும் அவை பெண்பால் என்றும் மற்றவை ஆண்பால் என்றும் குறிக்கிறது.

4.12.4.13. தொடர் கூறுகளுக்கு வெளியே உடன்பாடு (Inter-chunk agreement)

இந்தத் தொகுதியின் செயல்பாடு எழுவாய்-பனிலை உடன்பாடு பேன்ற தொடர் கூறுகளுக்கு வெளியே (inter-chunk) உள்ள உடன்பாட்டைக் கையாளுவது இந்தத் தொகுதியாகும். எழுவாயிலிருந்து பண்புக்கூறுகள் வினைச்சொல்லுக்கு நகலெடுக்கப்படுகின்றன. ஒருவர் தமது இலக்கு மொழிகளுக்கு தங்கள் சொந்த விதிகளை எழுத இயலும்.

4.12.4.14. தொடர் கூறுகளுக்கு உள்ளே உடன்பாடு (Intrachunk agreement)

இந்த தொகுதியின் செயல்பாடு தொடர் கூறுகளுக்கு இடையே உள்ள உடன்பாடுகளைக் (Intrachunk agreements) கையாளுவதாகும். பெயர்-பெயரடை உடன்பாடு போன்ற தொடர் கூறுகளுக்கு இடையே உள்ள உடன்பாடுகள் இந்தத் தொகுதியில் கையாளப்படுகின்றன. எடுத்துக்காட்டாக பெயர்ச் சொல்லின் பாலினத்தின் அடிப்படையில் பெயரடையின் பால் நிறுபப்படுகின்றது. ஒவ்வொரு மொழியும் செங்குத்தாக அவற்றின் இலக்கு மொழிகளுக்கு சொந்த விதிகளை எழுதும்.

4.12.4.15. TAM விபக்தி பிரிப்பான் (TAM Vibhakti Splitter)

இந்தத் தொகுதி 'விபக்தி' (வேறுமை உருபு / TAM) எடுக்கும் மற்றும் ஒவ்வொரு இடம்சார் சொல் குழுவிற்கும் அல்லது தொடர்கூறுக்கும் இதுபோன்ற பிற தகவல்களை எடுக்கும். தொடர்கூறில் அல்லது இடம்சார் சொல் குழுவில் செயல்பாட்டுச் சொற்களைச் (function words) சேர்க்கும்.

4.12.4.16. பிளவு விபக்தியில் உடன்பாட்டு விநியோகம்

இந்தத் தொகுதியின் செயல்ப்படு, விபக்தியிலிருந்து பண்புக்கூறுகளை அது பிரிக்கப்பட்ட சொற்களுக்கு விநியோகிப்பதாகும். பின்வரும் காரணத்தால் இந்தத் தொகுதி தேவைப்படுகிறது. பகுப்பாய்வு பகுதியில், இடம்சார் சொல் குழுவுவான் (word grouper) இடம்சார் சொல் குழுவில் உள்ள பல்வேறு கணுக்களிலிருந்து இடம்சார் சொல் குழுவிற்கான பண்புக்கூறுகளைக் கணக்கிடும். பின்னர் இடம்சார் சொல் பிரிப்பான் சொற்களைப் பிரிக்கும், ஆனால் சொற்கலுக்கு எந்த பண்புக்கூறுகளையும் தருவதில்லை. எனவே, இந்தத் தொகுதியின் முக்கியச் செயல்பாடு இடம்சார் சொல் குழுவின் பண்புக்கூறுகளை விநியோகிப்பதாகும். ஒருவர் தங்கள் குறிப்பிட்ட இலக்கு மொழிக்கு தங்கள் சொந்த விதிகளை எழுதலாம்.

4.12.4.17. இயல்புநிலை அம்சங்களை ஒதுக்கவும் (Assign Default Features)

பண்புக்கூறுகள் இல்லாத சொற்களுக்கு இயல்புநிலைப் பண்புகூறுகளை ஒதுக்குவதே இந்தத் தொகுதியின் செயல்பாடாகும். சில நேர்வுகளில் உருபு அனைத்து பண்புக்கூறுகளையும் கொடுக்காத சந்தர்ப்பங்கள் உள்ளன மற்றும் இடம்சார் சொல் பிரிப்பான் எல்லா பண்புக்கூறுகளையும் அது சேர்த்த செயல்பாடு சொற்களுக்கு வழங்காது. எனவே, இந்தத் தொகுதி இயல்புநிலை பண்புக்கூறுகளை வழங்குகிறது. சொல் உருவாக்கி (இந்த தொகுதிக்குப் பிறகு) சொல்லை உருவாக்க வேண்டி இது தேவை. ஒருவர் தங்கள் குறிப்பிட்ட இலக்கு மொழிக்கு தங்கள் சொந்த விதிகளை எழுதலாம்.

4.12.4.18. இருமொழிய அகராதி பொருத்தம் (Bilingual Mapping)

உருபனியல் பகுப்பாய்வால் அடையாளம் காணப்பட்ட வேர்ச்சொல் இலக்குமொழி நிகரன்களுக்கு வேண்டி இருமொழி அகராதியில் பார்க்கப்படும். இந்த இருமொழி அகராதி இலக்கு மொழியின் வேர்ச்சொல் நிகரன்களையும் அதன் வகைப்பாடு (பெயர், வினை போன்றவை) மேற்கோள் வாய்ப்பாடு மற்றும் பிற தேவையான தகவல்களைக் கொண்டிருக்கும். இந்த நிலை மூலமொழி இலக்கண மின்னுருபுகளுக்கு இலக்கு மொழி நிகரன்களை அடையாளம் காண அகராதியைப் பயன்படுத்துகிறது.

4.12.4.19. குறிப்பிட்ட இடம்சார் குழு பிரிப்பான்

இந்தப் பகுதி ஒரு சொல்லுக்கு விபக்தி மற்றும் பிற தகவல்களைக் கணிக்கும் மற்றும் அவற்றைத் தனிச்சொற்களாகத் திருப்பி அனுப்பும்.

4.12.4.20. உருபனியல் உருவாக்கம்/சொல் உருவாக்கம் (word generation)

ஒரு குறிப்பிட்ட இலக்கண வகைக்காக ஒரு சொல்லை உருவாக்க வேர்ச்சொல்லுடன் பொருத்தமான பின்னொட்டுகள் இணைக்கப்பட வேண்டும். உருவாக்கி மொழியின் உருபனியலுக்குத் தக்கவாறு வேர்ச்சொல்லை திரிபுற செய்யும் மற்றும் இலக்குமொழி சொல் வடிவை வெளியிடும். இவ்வாறு உருவாக்கப்பட்ட சொற்கள் முழு இலக்குமொழி வாக்கியத்தை உருவாக்க இணைக்கப்படும்.

4.12.4.21. வாக்கியநிலை உருவாக்கம்

இந்தச் செயல்பாடு தேவையானால் தொடர்பான்களை முதன்மையாக நிரல் மாற்றியமைக்கும். சில வாக்கியநிலை நோக்குகளும் இந்தப் பகுதியால் கையாளப்படும். எடுத்துக்காட்டாக பெயர், வினை உடன்பாடு போன்றவை இங்கு கையாளப்படும். இந்த மட்டம் தொடர்பகுதி மட்ட ஆய்வாகும். இடம்சார்சொல் வகுப்புகளைக் கொண்ட தொடர் பகுதிகளுக்கு பொருத்தமான சொற்கள் (அப்படியிருந்தால்) உருவாக்கப்படும். முதன்மை வினை, துணை வினை உடன்பாடு போன்ற உள் தொடர் பகுதி உடன்பாடு இங்கு காணப்படும்.

4.12.4.22. மதிப்பீடு செய்தல்

மதிப்பீடு செய்யும் இயக்க நுட்பம் தனி வாக்கியங்களுக்கும் உரைகளுக்கும் வேண்டி இறுதி பயன்பாட்டாளர்களுக்காக நிறுவப்படும். இது புரிந்து கொள்ளுதல் மற்றும் மொழிபெயர்ப்பு தரம் இவற்றிற்காக மதிப்பீடு செய்யப்படும்.

இரண்டாவது நிலையில் மதிப்பீடு ஒழுங்குமுறையின் வேகத்திற்காகவும் பயன்பாட்டிற்காகவும் செய்யப்படும்.

4.12.5. முடிவுரை

இத்திட்டம் இரண்டு கால கட்டங்களில் நடைபெற்றது. இத்திட்டத்தின் மொழி இணைகளில் ஒரு இணையான தமிழ்-மலையாளம், மலையாளம்-தமிழ் இயந்திரமொழிபெயர்ப்பு இராசேந்திரன் (திட்டத்தின் முதன்மை ஆய்வாளர்) மேற்பார்வையின் கீழ் தஞ்சாவூர் தமிழ்ப் பல்கலைக்கழகத்தில் செய்யப்பட்டது (2007-2011). இத்திட்டத்தின் கீழ் மேலே பட்டியலிடப்பட்டுள்ள எல்லா மொழி இணைகளுக்கும் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் உருவாக்கப்பட்டன. இவை யாவும் ஓரளவுக்கு இயந்திர மொழிபெயர்ர்ப்பு செய்யும் நிலையை எட்டியுள்ளன. குறிப்பாக இந்தி-பஞ்சாபி, இந்தி-உருது இணைகளின் இயந்திர

மொழிபெயர்ப்பு ஒழுங்கு முறைகள் நண்றாகச் செயல்படுகின்றன. கூகிள் மொழிபெயர்ப்பியின் வரவால் இந்திய அரசு இத்திட்டத்திற்கான நிதிநல்கையை நிறுத்திவிட்டது.

4.13. இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளின் ஒப்பீடு

இத்தலைப்பில் இந்திய இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகள் அவற்றின் அணுகுமுறைகள், மொழி இணைகள், பண்புக்கூறுகள் அடிப்படையில் ஒப்பிடப்பட்டு பட்டியலிடப்பட்டுள்ளன. திவேதி மற்றும் சுக்காதேவ் (Dwivedi and Sukhadeve 2010), சிதெந்தர் மற்றும் பாவா (Sitender and Bawa 2012), கர்ஜே மற்றும் கராதே Garje and Kharate 2013), படோதேகர் (Badodekar), நஸ்கர் மற்றும் பந்தோபாத்யா (Naskar and Bandyopadhya), ஆண்டோனி (Antony 2013), அதிதி கல்யாணி மற்றும் சஜ்ஜா Aditi Kalyani and Sajja 2015) (Godase and Govilka 2015) என்போர் இந்திய இயந்திர மொழிபெயர்ப்பு பற்றி எழுதிய கட்டுரைகளின் அடிப்படையில் கீழ்காணும் பட்டியல் உருவாக்கப்பட்டுள்ளது.

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள்	**அணுகுமுறை**	**மொழி இணைகள்**	**பண்புக்கூறுகள்**
அனுசாரக	நேரடி	பெங்காளி, கன்னடா, மராத்தி, பஞ்சாபி, தெலுங்கு இவற்றிலிருந்து இந்தி	பாணினிய இலக்கணதைப் பயன்படுதுகின்றது; மூலமொழிக்கும் இலக்குமொழிக்கும் இடையில் இடஞ்சார் சொற்களைப் பொருந்துகின்றது
இந்தியிலிருந்து பஞ்சாபி இயந்திரமொழி பெயர்ப்பு ஒழுங்குமுறை	நேரடி	இந்தியிலிருந்து பஞ்சாபி	சொல்லிலிருந்து சொல் நேரடி மொழிபெயர்ப்பு; உருபனியல் பகுப்பாய்வு, சொற்பொருள் மயக்கநீக்கம்,

மந்த்ரா	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து இந்தி, குஜராத்தி, தெலுகு; இந்தியிலிருந்து ஆங்கிலம், பெங்காளி, மராத்தி	கிளையமைப்பு இணைப்பு இலக்கண வடிவமைப்பு
மாத்ரா	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து இந்தி	மனித உதவியுடன் கூடிய மொழிபெயர்ப்பு; விதி அடிப்படையையும் பட்டறிவையும் பயன்பத்துகின்றது
சக்தி	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து இந்தி, மராத்தி, தெலுகு	விதி அடிப்படையையும் புள்ளியியல் அணுகுமுறையையும் இணைத்துச் செயல்படுகின்றது
அனுபாத்	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து பெங்காளி	கலப்பு ஒழுங்குமுறை; சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு என்-கிராம் அணுகுமுறையைப் பயன்படுத்துகின்றது; வாக்கிய நிலையில் வேலை செய்கின்றது
ஆங்கிலம்-கன்னடா இயந்திரமொழி பெயர்ப்பு ஒழுங்குமுறை	மாற்றல் அடிப்படை	ஆங்கிலத்திலிருந்து கன்னடா	உலகளாவிய எச்சத்தொடர் அமைப்பு இலக்கண வடிவமைப்பை பயன்படுத்துகின்றது.
சம்பர்க்	மாற்றல் அடிப்படை	பஞ்சாபி-இந்தி, தெலுங்கு-தமிழ்,	மொழிப் பகுப்பாய்விற்கு கணினிசார் பாணினி இலக்கண

		இந்தி-உருது, இந்தி-தெலுங்கு	அணுகுமுறையைப் பயன்படுத்துகின்றது; இதை இயந்திரம் கற்றலுடன் இணக்கின்றது
ஆங்கிலபாரதி	இடைமொழி சார்	ஆங்கிலம்-இந்தி	அடிக்கடி நேரிடும் பெயர்த்தொடர்களுக்கும் வினைத்தொடர்களுக்கும் மொழிபெயர்ப்பைப் பெறுவதற்காக விதி அடிப்படை, எடுத்துக்காட்டு அடிப்படை, புள்ளியியல் இவற்றைப் பயன்பத்துகிறது
உலகளாவிய இயற்கை மொழி ஆங்கிலம்-இந்தி மொழிபெயர்ப்பு ஒழுங்குமுறை	இடைமொழி சார்	ஆங்கிலம்-இந்தி	உலகளாவிய இயற்கை மொழியை இடைமொழியாகப் பயன்படுத்துகின்றது
ஆங்கிலம்-இந்தி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	புள்ளியியல் சார் இயந்திர மொழிபெயர்ப்பு	ஆங்கிலம்-இந்தி	விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பையும் தொடர் அடிப்படையிலான புள்ளியியல் இயந்திர மொழிபெயர்ப்பையும் இணைக்கின்றதும்
ஆங்கிலம்-மலையாளம் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	புள்ளியியல் சார் இயந்திர மொழிபெயர்ப்பு	ஆங்கிலம்-மலையாளம்	ஒருமொழிய மலையளத் தரவுத்தொகுதியையும் புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு இருமொழிய ஆங்கிலம்/மலையாளம்

			தரவுத்தொகுதியையும் பயன்படுத்துகின்றது.
வாசானுபாத	எடுத்துக்காட்டுசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	இருமொழிய பெங்காளி அசாமீஸ்	முன் செயற்பாங்கும் பின்செயற்பாங்கு செயல்பாடு, நீண்ட வாக்கியங்கள் நிறுத்தற்குறிகள் வரும் இடத்தில் துண்டுபடுத்தப்படும்; பொருந்தாத விளைவுகளுக்குப் பின்தடம்பற்றும்
அனுபாரதி	எடுத்துக்காட்டுசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	ஆங்கிலம்-இந்தி	கலப்பு எடுத்துக்காட்டு அடிப்படையிலான ஒழுங்குமுறையான இது அமைப்பொழுங்கு அடிப்படையையும் எடுத்துக்காட்டு அடிப்படை அணுகுமுறையையும் தன்னுடன் இணைத்துக்கொள்ளும்.
சிவா	எடுத்துக்காட்டுசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	ஆங்கிலம்-இந்தி	மொழியியல் தகவல்களை ஊகிப்பதற்காக மொழியியல் விதிகளையும் புள்ளியியல் அணுகுமுறையும் பயன்படுத்தும்.

4.14. தமிழ் சார் இயந்திர மொழிபெயர்ப்பு

தமிழ்சார் இயந்திர மொழிபெயர்ப்புகள் பல மேற்கொள்ளப்பட்டு தமிழ் நாட்டில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சிக்கு வித்திட்டன.

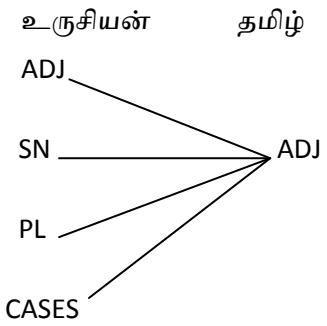
4.14.1. தமிழ் சார் இயந்திர மொழிபெயர்ப்பு முயற்சிகள்

இங்கு தரப்பட்டுள்ள இயந்திர மொழிபெயர்ப்பு திட்டங்கள் பெரும்மாலும் பிற மொழிகளை தமிழுக்கு இயந்திர மொழிபெயர்ப்பு செய்ய முயற்சித்த செயல்பாடுகளை உட்படுத்தும். முன்னர் குறிப்பிட்ட இந்தியா அளவிலான மொழிபெயர்ப்புகளிலும் தமிழ் இலக்குமொழியாகவும் மூல மொழியாகவும் செயலாற்றுகின்றது.

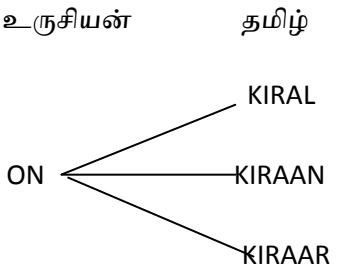
4.14.1.1 தமிழ் உருஷ்யன் மொழிபெயர்ப்புத்திட்டம்

1985-ஆம் தமிழ் பல்கலைக்கழகத்தில் (Tumtst Tamil University Machine Inaslation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது (Chellammuthu et al 1984, Chellemuthu 2002). இதன் மூலம் ருஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிப்பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக்கழகம் கண்ணி அறிவியல்துறை, மொழியியல் துறை மற்றும் மொழிப் பொயர்ப்புத் துறை ஆகியவைகளின் கூட்டு முயற்சியில் உருவானதாகும்.

உருசியன் தமிழ் மொழிபெயர்ப்பிற்கான படிநிலைகளின் விளக்கம் கீழே தரப்பட்டுள்ளன. மூலமொழி உருசியனை L1 என்றும் தமிழ் மொழியை L2 என்றும் குறிப்பிடுகின்றன. முதற்கட்டமாக மூலமொழியில் உள்ள ஒரு செய்தி ஒலிபெயர்ப்பு செய்யப்பட்டு கணிப்பொறிக்கு உள்ளீடு செய்கின்றன. இரண்டாம் கட்ட நிலையில் மூல மொழியில் உள்ள செய்தி பகுப்பாய்வு செய்யப்படுகிறது. இந்நிலையில் சொற்களின் இலக்கண கூறுகள் கண்டறியப்படுகின்றன. இதற்கு ருஷ்யன் - தமிழ் அகராதிகள் உதவுகிறது. இப்பகுப்பாய்வு மூலம் ருஷ்ய மொழியில் உள்ள செய்திகள் NP, VP, NN, ADF என பாகுபடுத்தப்படுகின்றன. மூன்றாம் கட்ட நிலையில் பகுப்பாய்வு செய்யப்பட்டுள்ள மூலமொழியின் வார்த்தையில் உள்ள சொற்களின் படிவம், வாக்கிய அமைப்பு போன்றவைகளுக்கு இலக்கு மொழியில் நிகராக உள்ள சொற்களின் வடிவம், வாக்கிய அமைப்பு ஆகியன மாற்றம் (Transfer) என்ற அமைப்பின் மூலம் கண்டறியப்பட்டு மொழி மாற்றம் செய்யப்படுகின்றன. இந்நிலையில் இவ்விரு மொழிகளுகட்கும் இடையே உள்ள முரண்பாட்டு அறிவு (Contrast Knoweldge) பயன்படுத்தப்படுகின்றன. கருத்தாடலைப் பொருத்தவரை இவ்விரு மொழிகளுக்கு இடையே ஒன்றுக்கு பல (One-to-Many) உறவு உள்ளது. பெயரடைச்சொல்லை பொறுத்தவரையில் இவ்விரு மொழிகளுக்கு இடையே (many-to-One) உறவு உள்ளது. இவ்வுறவினை கீழ்கண்ட படம் விளக்குகிறது.



மேலும் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமை உருபுகள் ஒன்றுகுப் பல (one-to-many) என்ற உறவைக்காட்டுகின்றன.



இதுபோன்ற இவ்விரு மொழிகளுக்கு இடையே உள்ள வேறுபாடுகள் விதிமுறைகளாகவும் வழிமுறைகளாகவும் மாற்றம் ((Transfer)) என்ற அமைப்பின் கீழ் உள்ளன. பகுப்பாய்வு செய்யப்பட்ட மூலமொழி செய்திகள் மாற்ற (Transfer) அமைப்பில் உள்ளீடாகச் செல்கிறது. இந்நிலையில் மூலமொழி அமைப்பிற்கு நிகரான இலக்கு மொழி அமைப்பிற்குச் செய்தி மொழிமாற்றம் செய்யப்படுகிறது. நான்காவது கட்ட நிலையில் மொழி மாற்றம் செய்யப்பட்டுள்ள இலக்கு மொழி கட்டமைப்புகள் இலக்கு மொழியின் தன்மைக்கேற்ப அமைவதில்லை. இவ்வாறு கிடைக்கப்பெற்ற இடைநிலை மொழிமாற்ற அமைப்புகளை இலக்கு மொழியில் தன்மைக்கேற்ப உருவாக்கித் தருவது தான் உருவாக்கம் (Genaration) என்ற பகுதியின் பணியாகும். இந்நிலையில் மொழிமாற்றம் செய்யப்பட்ட செய்திகள் உள்ளன. இச்செய்தியை மூலமொழியின் வரிவடிவத்திலே அல்லது ஒலிபெயாப்பு வடிவத்திலே தரும் பணியை வெளியீடு (Output) என்ற அமைப்பு தருகிறது. மொழிபெயர்ப்பு செய்யப்பட்ட செய்தியைக் காட்சித் திரையில் காணலாம். தேவைப்படின் அச்சுப்பொறி வழி அச்சிட்டுக்கொள்ளலாம். உருசியன்-தமிழ் மொழிபெயர்ப்பைப்

பொருத்தவரையில் வெளியீட்டை தமிழ் எழுத்துக்களில் காண்பதற்கும், அச்சிட்டுக் கொள்வதற்கும் வசதிகள் உள்ளன. அதுபோல் ஒலிபெயர்ப்பு முறையிலும் விடையைப் பெறலாம்.

4.14.1.2. இந்தி-தமிழ் மொழிபெயப்புக்கு அனுசாரகா ஒழுங்குமுறை

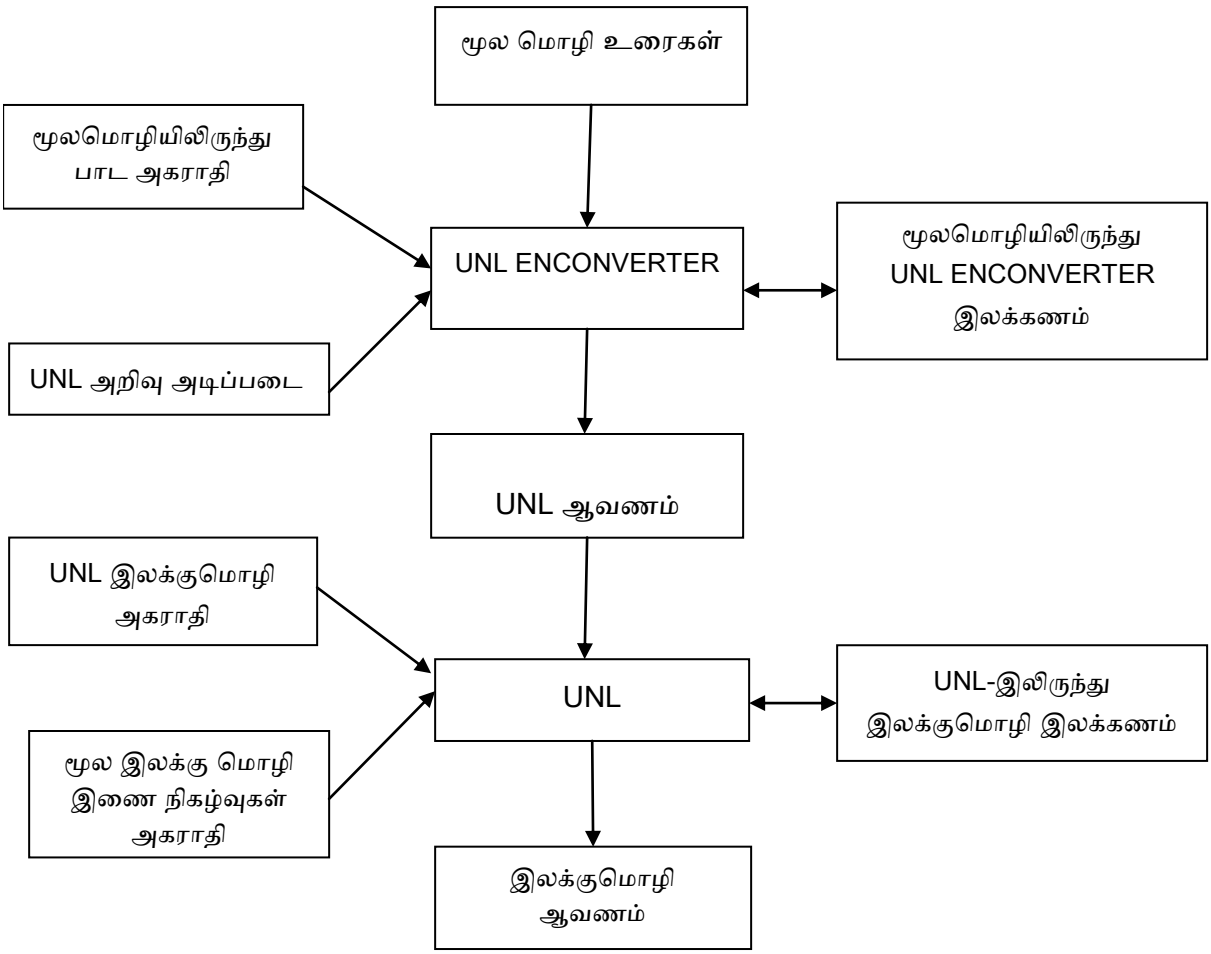
அனுசாரகா சட்டகத்தின் செயல்முறையை அடிப்படையாகக் கொண்ட இந்தி-தமிழ் மொழிபெயர்ப்பிற்கான அனுசாரகா இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை ஐ.ஐ.டி கான்பூரில் 1990களில் ராஜீவ் சங்கல், வினித் சைத்தன்யா, அம்பாகுல்கர்னி ஆகியோரின் மேற்பார்வையில் இராசேந்திரனால் மேற்கொள்ளப்பட்டது. இம்மொழிபெயர்ப்பு ஒழுங்குமுறை இந்திய மொழிகளுக்கெல்லாம் பொதுவான ஒரு மொழிபெயர்புச் சட்டத்தின் அடிப்படையில் உருவாக்கபட்டது. ராஜேந்திரன் இதன் உருவாக்கதில் பெரும் பங்காற்றினார். ஓரளவுக்கு வெற்றியும் கிடைத்தது. இந்திய இயந்திரமொழிபெயர்ப்பின் தொடக்க காலகட்டத்தில் இதன் மொழிபெயர்ப்பு வெளியீடு அதிக ஆர்வத்தைத் தந்தது. இந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் அடுத்தகட்ட உருவாக்கம் சென்னையில் உள்ள AUKBCRCஆல் மேற்கொள்ளப்பட்டு இராசேந்திரனின் மேற்பார்வையில் முடிக்கப்பட்டது. தமிழுக்கு முதல்முதலாக ஒரு இயந்திரமொழிபெயர்ப்பு கருவி உருவாக்குவதற்கும் உருபனியல் பகுப்பாய்வி உருவாக்குவதற்கும் இது ஒரு வாய்ப்பாக அமைந்தது. அனுசாரகா தான் பின்னர் உருவாக்கப்பட்ட மேம்பட்ட இந்திய இயந்திரமொழிபெயர்ப்புத் திட்டங்களுக்கு அடித்தளமாக அமைந்தது. இது சொல் அடிப்படையிலான நேரடி அணுகுமுறை தழுவிய இயந்திர மொழிபெயர்ப்புக் கருவியாகும். இதில் தொடரியல் பகுப்பாய்வு பயன்படுத்தப்படவில்லை. இந்திய மொழிகள் யாவும் ஒரே தொடரப்பைப் பங்கிட்டுக்கொள்வதால் தொடரியல் பகுப்பாய்வுக்குத் தேவை இல்லை என்பது இதன் வாதம். சொற்களுக்கும் திரிபு அலகுகளுக்கும் அகராதி உருவாக்கப்பட்டது. துணைவினைக் குழுமங்களுக்கும் ஒன்றுக்கொன்றான அகராதி உருவாக்கப்பட்டது. இவற்றைப் பயன்படுத்தித்தான் இந்தி-தமிழ் அனுசாரகா உருவாக்கப்பட்டது. (அனுசாரகா என்றாலே உதவும் கருவி என்று அர்த்தம்.

4.14.1.3. உலக வலைப்பின்னல் மொழி – தமிழுக்கான இடைமொழி இயந்திர மொழிப்பெயர்ப்பு

இந்த அணுகுமுறை சென்னையில் உள்ள அண்ணா பல்கலைக்கழகத்தில் கணிப்பொறி மொழியியல் துறையில் மேற்கொள்ளப்பட்டது. (Dhanabalan & Geetha, 2004). இந்த அணுகு முறையில் இடையீட்டு மொழியின் உருப்படுத்தம் பயன்படுத்தப்படுகிறது. மூலமொழி அகராதியையும் இலக்கணத் தகவல்களையும் பயன்படுத்தி இடைமொழி உருப்படுத்தம்

மூலமொழியிலிருந்து இலக்கு மொழிக்கு மொழிப்பெயர்ப்பு செய்ய பயன்படுத்தப்படும் இடைமொழி அமைப்பு உலகளாவிய வலைப்பின்னல் மொழி (Universal Networking Language (UNL)) ஆகும். இம்மொழி இடைமொழி அமைப்பாகத் திட்டமிடப்பட்டு மூலமொழியிலிருந்து இலக்கு மொழிக்கும் இலக்கு மொழியிலிருந்து மூலமொழிக்கும் தானியங்கி இயந்திர மொழிபெயர்ப்புயைச் சாத்தியமாக்குகின்றது. UNL-ஐ பயன்படுத்திய மொழிபெயர்ப்பு ஒழுங்குமுறை மூலமொழியிலிருந்து UNL-க்கு மாற Encoverter-ஐயும் UNL-இலிருந்து இலக்கு மொழிக்கு மாற Deconverter பயன்படுத்துகின்றது.

இந்த அணுகுமுறையின் திட்டவடிவம் கீழே தரப்பட்டுள்ளது.



மேற்கண்டவாறு படத்தில் காட்டியவாறு UNL அமைப்பு வேலை செய்து UNL மொழியிலிருந்து தமிழுக்கு மொழிபெயர்ப்பு செய்யும்.

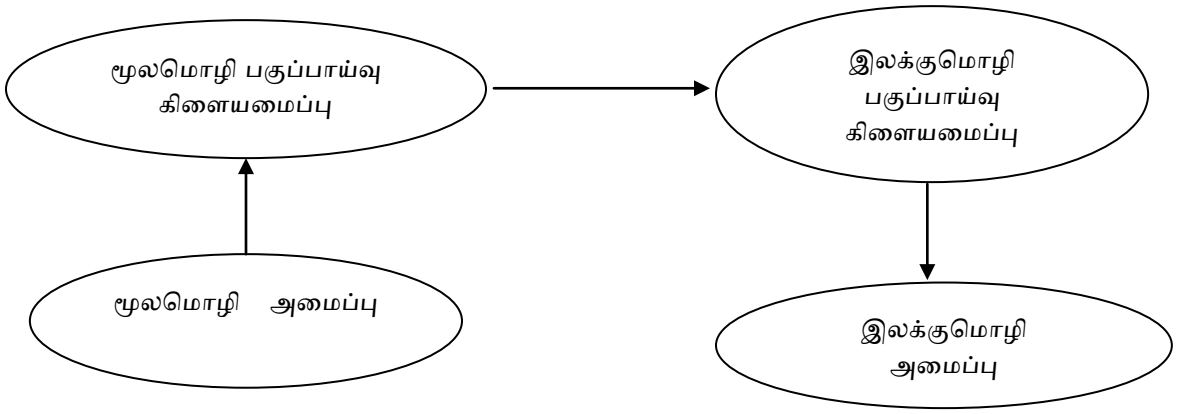
4.14.1.4. ஆங்கிலத்திலிருந்து மொழியியல் நூல்களைத் தமிழில் மொழிப்பெயர்க்கும் திட்டம்

ஆங்கிலத்திலிருந்து மொழியியல் நூல்களைத் தமிழில் மொழிபெயர்க்க ஒரு இயந்திர மொழி பெயர்ப்புக் கருவி உருவாக்க ராஜேந்திரன் ஒரு திட்டத்தை உருவாக்கி செயல்படுத்த முயன்றார். தமிழில் முதுகலை மொழியியல் பாடத்திட்டத்திற்கு வேண்டிய மொழியியல் நூல்களைத் தமிழில் மொழிப்பெயர்க்க வேண்டிய கட்டாயம் ஏற்பட்டதால் இத்தகைய ஆய்வு மேற்கொள்ளப்பட்டது. இவ்வாய்வு மொழியியல் நூல்களை ஆங்கிலத்திலிருந்து தமிழுக்கு

மொழிபெயர்க்க உதவும் ஒரு கருவி தயாரிப்பதை நோக்கமாக கொண்டது. இக்கருவி உருவாக்கம் பின்வரும் நான்கு பகுதிகளைக் கொண்டது.

1. ஆங்கில மொழி உரைப் பகுப்பாய்வு
2. ஆங்கில-தமிழ் மொழிபெயர்ப்பு அகராதி
3. ஆங்கில – தமிழ்மொழி மாற்றல் விதிகள்
4. தமிழ் மொழி உரை உருவாக்கம்

பின்வரும் படம் இந்த மொழிப்பெயர்ப்புக் கருவியின் செயல்பாட்டை விளக்கும்.



மாற்றல் அடிப்படையிலான இந்த இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கம் பின்வரும் செயல்பாடுகளை உட்படுத்தும்.

உருபனியல் பகுப்பாய்வு: உள்ளீட்டு உரையின் புற வடிவங்கள் சொல்வகைப்பாடுகளாகவும் (எ.கா. பெயர்ச்சொல், வினைச்சொல், முதலியன) மற்றும் துணை வகைப்பாடுகளாகவும் (வேறுமை, எண், பாலினம், காலம் போன்றவை) வகைப்படுத்தப்படுகின்றன. ஒவ்வொரு புற வடிவத்திற்கும் சாத்தியமான "பகுப்பாய்வுகள்" அனைத்தும் இந்தக் கட்டத்தில் சொல்லின் சொல்லனுடன் (lemma) வெளியீடு செய்யப்பட்டன.

சொல் வகைப்படுத்தல்: எந்தவொரு உரையிலும் சில சொற்களுக்கு ஒன்றுக்கு மேற்பட்ட அர்த்தங்கள் இருக்கலாம், இது பகுப்பாய்வில் பொருண்மை மயக்கத்தை ஏற்படுத்துகிறது. உள்ளீட்டின் சூழலில் சரியான அர்த்தத்தைத் தீர்மானிக்க முயற்சிக்க சொல் வகைப்படுத்தல் மேற்கொள்ளப்பட்டது. சொல்வகைப்படுத்தல் ஒரு சொல்லின் சூழலைப் பார்த்து சொல்

வகைப்படுத்தலைச் செய்தது. இது சொல்வகைப்பாடு அடையாளப் படுத்தலையும் சொற்பொருண்மை மயக்கநீக்கத்தையும் உள்ளடக்கியது.

சொல் மாற்றல்: இது அடிப்படையில் அகராதி மொழிபெயர்ப்பு ஆகும்; மூலமொழிச் சொல்லன் (ஒருவேளை அர்த்தம் பற்றிய தகவலுடன்) இருமொழி அகராதியில் பார்க்கப்பட்டு மொழிபெயர்ப்பு நிகரன் தேர்வு செய்யப்பட்டது.

கட்டமைப்பு மாற்றல்: முந்தைய கட்டங்கள் சொற்களைக் கையாளும் போது, இந்த நிலை பெரிய கூறுகளைக் கையாண்டது; எடுத்துக்காட்டாக, சொற்றொடர்கள் மற்றும் இடைச்சொல்கள். இந்தக் கட்டத்தின் பொதுவான அம்சங்களில் பாலினம் மற்றும் எண்ணின் ஒத்திசைவு மற்றும் சொற்கள் அல்லது சொற்றொடர்களை மறு வரிசைப்படுத்துதல் ஆகியவை இதில் அடங்கும்.

உருபனியல் உருவாக்கம்: கட்டமைப்பு மாற்றல் கட்டத்தின் வெளியீட்டிலிருந்து இலக்கு மொழி புற வடிவங்கள் உருவாக்கப்பட்டன.

இம்மொழிபெயர்ப்பு ஆங்கில மொழியியல் நூல்களை தமிழுக்கு மொழிபெயர்பதைக் குறிக்கோளாக்கக் கொள்வதால் இதை ஒரு துணைமொழி அணுகுமுறை (sub-language approach) அடிப்படையிலான மொழிபெயர்ப்பு என்று கூறலாம். மேலும் இதன் நோக்கம் தொடரியல் நூல்களை மொழிபெயர்ப்பதை குறிப்பாக சாம்ஸ்கியின் Aspects of the theory of syntax என்ற நூலை மொழிபெயர்ப்பதில் கவனக்குவிப்பு செய்ததால் இது ஒரு கட்டுப்படுத்தப்பட இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையாக அமைந்தது. பொதுவான ஆங்கில மொழியை மொழிபெயர்ப்பதால் ஏற்படும் பொருண்மைமயக்கம் அடிப்படையிலான சிக்கல்கள் இம்மொழிபெயர்ப்பில் குறைவு. இருமொழிய மொழிபெயர்ப்பு அகராதியும் மேற்சொன்ன சாம்ஸ்கியின் நூலுக்காக உருவாக்கப்பட்டது. மாதிரிக்காக சாம்ஸ்கியின் நூலிலிருந்து வாக்கியங்கள் தேர்ந்தெடுக்கப்பட்டு அவை எளிய வாக்கியங்களாக மனித முயற்சியால் உடைக்கப்பட்டன. இவ்வாக்கியங்கள் டோக்கனைஸ் செய்யப்பட்டு, சொல்வகைப்பாட்டிற்காக அடையாளப்படுத்தப்பட்டு சாம்ஸ்கியின் சூழல்கட்டுப்பாடில்லாத இலக்கண வடிவமைப் பயன்படுத்தி தொடரியல் பகுப்பாய்வு செய்யப்பட்டது.

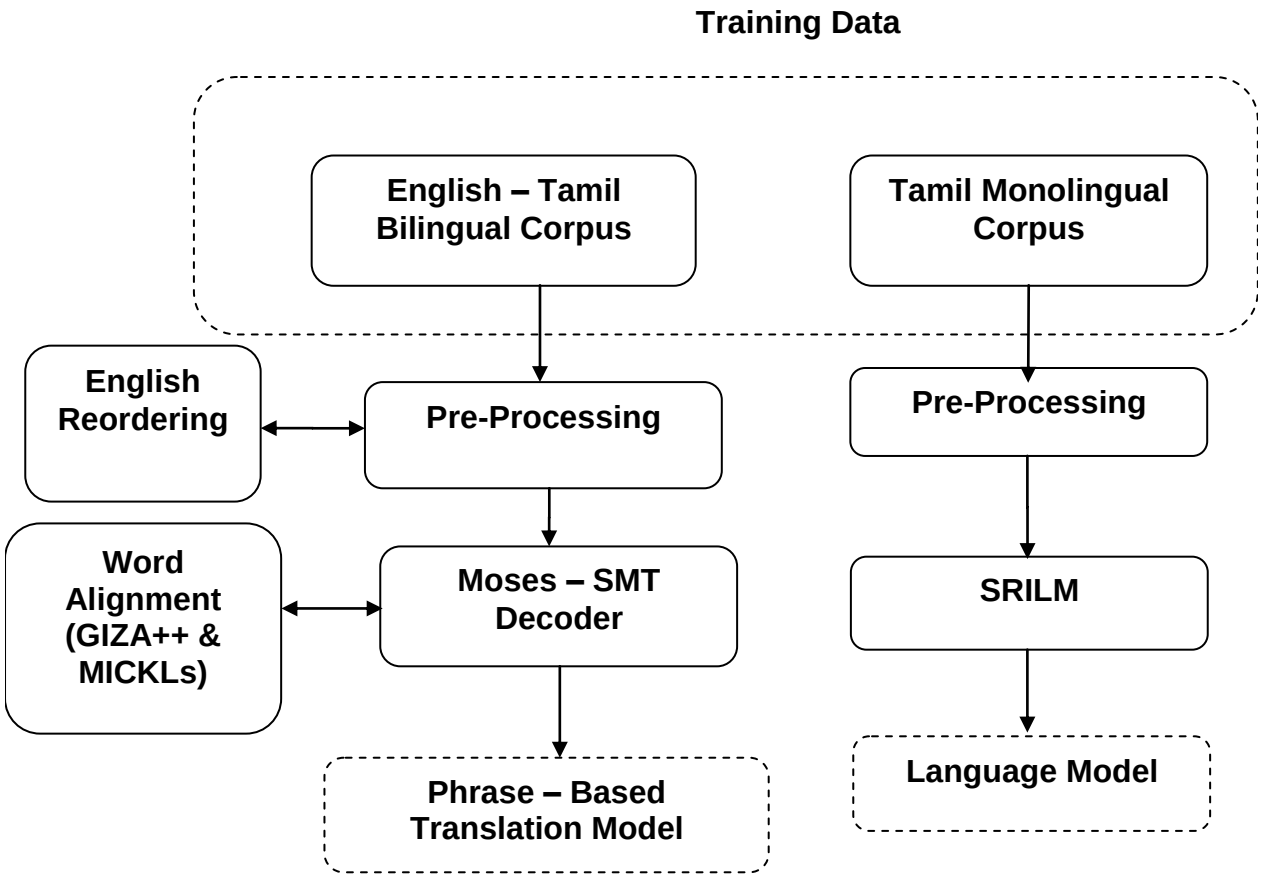
ஏற்கனவே மேற்கொண்ட ஆங்கிலம்-தமிழ் சொல்வரிசை முரண்பாட்டு ஆய்வின் அடிப்படையில் ஆங்கிலத்திற்கும் தமிழுக்கும் இடையிலான தொடரியல் சார் வேறுபாடுகளை அறிந்துகொண்டு மாற்றல் விதிகள் எழுதப்பட்டன. தொடரியல் பகுப்பாய்வில் விளையும்

வெளியீடு இம்மாற்றல் விதிகளைப் பயன்படுத்தி தமிழ்மொழியின் தொடரியல் அமைபுக்கு மாற்றப்பட்டது. உருபனியல் பகுப்பாய்வி தொடரமைபுக் கிளையின் இறுதியில் வரும் சொற்களை உருபனியல் விதிகளைப் பயன்படுத்தி உருபன்களாகப் பிரித்தது (சொல்லனாகவும் ஒட்டுக்களாகவும்). சொல்லன்கள் ஆங்கிலம்-தமிழ் இருமொழிய மொழிபெயர்ப்பு அகராதியில் பார்க்கப்பட்டு அவற்றிற்கு இணையான தமிழ் சொல்லன்கள் தெரிந்தெடுக்கப்பட்டன. ஒட்டுக்கள் உருபனியல் பண்புக்கூறுகளாக ஆயப்பட்டு (பெயர், பின்னுருபு, வினை, காலம்) ஆங்கில உருபனியல் பண்புக்கூறு-தமிழ்உருபனியல் பண்புக்கூறு பொருத்த அகராதியைப் பயன்படுத்தி மூலமொழி உருபனியல் பண்புக்கூறுகளுக்கு இணையான தமிழ் உருபனியல் பண்புக்கூறுகள் கண்டுபிடிக்கப்பட்டன. இதன் வெளியீடு உருபனியல் அல்லது சொல் உருவக்கிக்கு அனுப்பபட்டு தமிழ் உரையாக மாற்றப்படது. இவ்விளக்கங்கள் காமாட்சியும் இராசேந்திரனும் (Kamakshi and Rajendran) வெளியிட்ட Preliminaries to the Preparation of a Machine Aid to Translate Lingustic Texts in English into Tamil என்ற நூலில் விரிவாகத் தரப்பட்டுள்ளது.

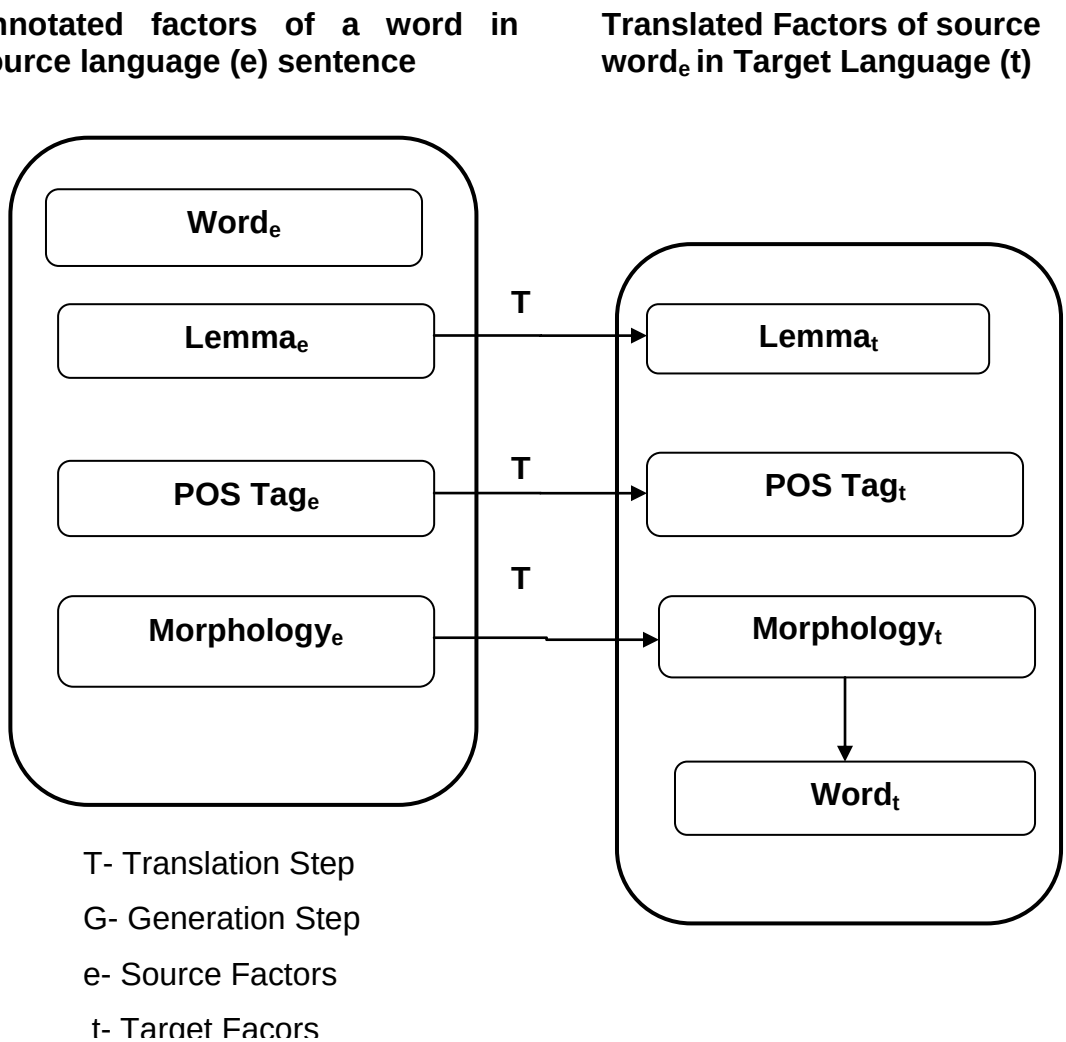
4.14.1.5. இணைத் தரவுத்தொகுதியைப் பயன்படுத்தி ஆங்கிலம்-தமிழ் இயந்திரமொழிபெயப்பு ஒழுங்குமுறை

இணைத் தரவுத்தொகுதியைப் பயன்படுத்தி ஆங்கில-தமிழ் இயந்திரமொழிபெயப்பு ஒழுங்குமுறை (English-Tamil Machine Translation system using parallel corpus) என்ற ஆய்வுத்திட்டம் 2010இல் இராசேந்திரனால் தொடங்கப்பட்டு 2013இல் முடிவுற்றது. இது பற்றிய விளக்கம் நூல் வடிவில் வெளியிடப்பட்டுள்ளது (Rajendran and Vasuki 2019). இதற்கென்று ஆங்கில-தமிழ் இணை இருமொழியத் தரவுத்தொகுதியும் தமிழ் ஒருமொழியத் தரவுத்தொகுதியும் உருவாக்கப்பட்டன. இணைத் தரவுத்தொகுதியில் வாக்கியங்கள் இணைகளாக வரிசைபடுத்தப்பட்டன. இது தொடர் அடிப்படையிலான புள்ளியியல்சார் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையின் கண்படும். மூலமொழி வாக்கியங்கள் சொல் வகைப்பாடுக்கு அடையாளப்படுத்தப்பட்டன. மேலும் தொடர்கூறுகளுக்காகவும் பகுக்கப்பட்டன. பின்னர் தேவையான மெபொருள்களைப் பயன்படுத்தி இயந்திர மொழிபெயர்ப்பு செய்யப்பட்டது. விளைவு திருப்திகரமாக இருந்தது. இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையின் கட்டமைப்புப் படம் கீழே தரப்பட்டுள்ளது.

Proposed System Architecture



பின்வரும் பிளாக் வரைபடம் மொழிபெயர்ப்புச் செயல்பாட்டையும் உருவாக்கத்தையும் காட்டும்.



இணையான தரவுத்தொகுதியைப் பயன்படுத்துவதன் மூலம் ஆங்கிலத்திலிருந்து தமிழ் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு புதிய முயற்சி. அமைப்பின் துல்லியம் மொழிகளில் கிடைக்கும் இணையான தரவுட்தொகுதியின் அளவைப் பொறுத்தது. உருனியல் தகவல், சொல்வகப்பாட்டுக்கு வகைப்படுத்தல் போன்ற மொழியியல் செய்திகளைச் சேர்ப்பது அமைப்பின் துல்லியத்தை மேம்படுத்தும். இது காரணி முறை (factored method) என்று அழைக்கப்படுகிறது. தற்போது இந்த அமைப்பு அதன் அடிப்படை கட்டத்தில் மட்டுமே உள்ளது. இதனால் எளிய

வாக்கியங்களை மட்டுமே மொழிபெயர்க்க முடியும். தரவுத்தொகுதி அளவை அதிகரிப்பதன் மூலமும் மொழியியல் தகவல்களை உள்ளிடுவதன் மூலமும் பெருக்குதல் இந்த ஒழுங்குமுறையை மேம்படுத்தலாம். இவ்வியந்திர மொழிபெயர்ப்பு பற்றிய விரிவான விளக்கம் இரசேந்திரன், வசுகி (Rajendran and Vasuki 2019) வெளியிட்ட "English-Tamil Machine Translation system using parallel corpus" என்ற நூலில் விவாக விளக்கப்பட்டுள்ளது.

4.14.1.6. தமிழுக்கான புள்ளியியல் சார் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகள்

அமிர்தா பல்கலைக் கழகத்தைச் சார்ந்த ஆனந்த் குமார் (Anand Kumar 2015) ஆங்கிலம்-தமிழ் கருத்துப் பரிமாற்றத்திற்கான புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை (Stastical Machine Translation System) உருவாக்கினார். மேலும் ஆங்கிலத்திலிருந்து தமிழ் மொழிக்கு காரணியாலான புள்ளியியல் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைய (Factored Statistical Machine Translaton System) முன்மொழிந்தார் (Anand Kumar et al 2010).

4.14.1.7. இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் அமிர்தா விஷ்வ வித்யபீடத்தின் பங்களிப்புகள்

கோயம்புத்தூர் அமிர்தா பல்கலைக்கழகம் தமிழ் சார்ந்த இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதற்கான பல திட்டங்களைக் கைக்கொண்டு செயலாற்றியது. இது CDAC, புனேயைச் சார்ந்திருந்த தர்பாரியின் தலைமையில் நடைபெற்ற EILMT (English to Indian Languge Machine Translation) கூட்டமைப்பு திட்டத்தின் (consortia project) கீழ் ஆங்கிலத்திலிருந்து தமிழுக்கு இயந்திர மொழிபெயர்ப்பை மேற்கொண்டு முடித்தளித்தது. இது TAG இலக்கணத்தை அடிப்படையாகக் கொண்டது. அமிர்தா பல்கலைக்கழகத்தைச் சார்ந்த கேபி சோமன் ஆங்கிலம்-தமிழ் மொழிபெயர்ப்புத் திட்டப் பகுதியின் முதன்மை ஆய்வாளராகச் செயலாற்றினார். இத்திட்டதின் வடிவமைப்பு விஜய் கிருஷ்ணமேனோனால் செய்யப்பட்டது. அமிர்தா பல்கலைக்கழகக் கணிசார் பொறியியல் மற்றும் வலைப்பின்னல் மையத்தைச் (Centre for Computational Engineering and Networking (CEN) சார்ந்த/சார்ந்திருந்த (கேபி சோமன், கோவிந்த மேனோன், லோகநாதன், சரவணன், விஜய் கிருஷ்ண மேனோன், ஆனந்குமார், தனலெட்சுமி, இராசேந்திரன் ஆகியோரைக் கொண்ட) இயற்கை மொழி ஆய்வுக் குழு (NLP Research Group) இவ்வாய்வுத்திட்ட வெற்றிக்காக முழுமையாகப் பாடுபட்டனர். இதன் இயந்திர மொழிபெயர்ப்பு வெளியீடு (விடுவரல்) கூறத்தக்கத் தரத்தில் அமைந்தது.

ஆனந்த் குமார் (Anand Kumar 2015) தனது முனைவர் பட்ட ஆய்வின் ஒரு பகுதியாகக் கணினி மொழியியலின் அடிப்படையில் ஆங்கிலம்-தமிழ் கருத்துப் பரிமாற்றத்திற்கான புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை (Stastical Machine Translation System) உருவாக்கினார். மேலும் ஆங்கிலத்திலிருந்து தமிழ் மொழிக்கு காரணியாலான புள்ளியியல் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைய (Factored Statistical Machine Translaton System) முன்மொழிந்தார் (Anand Kumar et al 2010). மேலும் தமிழ் இணையக் கல்விநிறுவனத்தின் நிதியுதவியுடன் "தமிழ் கற்பிப்பதற்கான கணினிக் கருவிகள்" (Computational Tools for Teaching Tamil) திட்டத்தின் கீழ், அமிர்தா பல்கலைக்கழகம் ஆங்கிலம்-தமிழ் இயந்திர மொழிபெயர்ப்பிற்கான விதி அடிப்படையிலான மொழிபெயர்ப்பு ஒழுங்கமை உருவாக்கியுள்ளது (Rajendran and Anand Kumar 2018). இதில் ஆங்கிலத்தின் ஸ்டான்போர்ட் பல்கலைக்கழக சார்பு பகுப்பாய்வியின் வெளியீடு தமிழ் சார்பு பகுப்பாய்வு வெளியீடாக மாற்றப்பட்டு மற்றும் உருபனியல்-தொடரியல் உருவாக்கியைப் பயன்படுத்தி தமிழ் உரை உருவாக்கப்படுகிறது. இந்த அமைப்பு அதன் தொடக்க நிலையில் உள்ளது. அது மேலும் மேம்படுத்தப்பட வேண்டும்.

4.14.1.8. இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் AUKBCRC இன் பங்களிப்பு

இந்தி-தமிழ் இயந்திரமொழிபெயர்ப்பு உதவிக்கருவி உருவாக்க ஐஐடி கான்பூரில் தொடங்கப்பட்ட அனுசரகா திட்டத்தை AUKBCRC நிறைவு செய்தது. இது இராசேந்திரனின் மேற்பார்வையின் கீழ் நடைபெற்றது. இந்தியமொழியிலிருந்து இந்தியமொழி இயந்திரமொழிபெயர்ப்பு (ILILMT) கூட்டமைப்பு திட்டத்தின் (consortium project) கீழ் இந்தி-தமிழ், தமிழ்-இந்தி இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பை உருவாகியது. இது தவிர, AUKBCRC ஆங்கிலம்-தமிழ் மொழிபெயர்ப்பிற்கான ஒரு இயந்திரமொழிபெயர்ப்பு ஒழுங்கமைப்பையும் உருவாக்கியுள்ளது. இது ஒரு விதி அடிப்படையிலான ஒழுங்கமைப்பு ஆகும். இது இராசேந்திரனின் மேற்பார்வையில் நடைபெற்றது.

4.14.1.9. இயந்திர மொழிபெயர்ப்பை மேம்படுத்துவதில் தமிழ் பல்கலைக்கழகத்தின் பங்களிப்பு

தமிழ்ப் பல்கலைக்கழம் தான் முதன் முதலில் ஒரு ஒழுங்கான இயந்திர மொழிபெயர்பில் ஈடுபட்டது. இதற்கு முக்கிய காரணமாக இருந்தவர் அன்றைய துணைவேந்தர் பேராசிரியர் வி.ஐ. சுப்பிரமணியம் ஆவார். இவர் மேற்பார்வையில் ரஷ்ய-தமிழ் மொழிபெயர்ப்பு முயற்சி மேற்கொள்ளப்பட்டது (Chellamuthu et al 1984). மேலும் ஆங்கில மொழியியல் பாட நூல்களை

தமிழில் மொழிபெயர்க்க ஒரு இயந்திர உதவிக்கருவியைத் தயாரிக்க முயற்சி எடுக்கப்பட்டது (Rajendran and Kamashi 2004).

4.14.2. தமிழ்சார் மொழிபெயர்ப்பு நடவடிக்கைகள்

4.14.2.1 உருபனியல் ஆய்வு (Morphological Analysis)

இது வேர் சொற்களைக் கண்டுபிடிப்பதையும் அதன் இலக்கண பண்புப் கூறுகளைக் கண்டுபிடிப்பதையும் உள்ளடக்கும். தமிழ் ஒரு திரிபு (Inflection) வளமுள்ள மொழி ஆகையால் அதன் பகுப்பாய்வு ஆங்கிலத்தின் உருபனியல் பகுப்பாய்வை விடச் சிக்கல் வாய்ந்ததாக அமையும். புணர்ச்சி விதிகள் தமிழில் உருபனியல் ஆய்வி உருவாக்குவதை கடினமாக்குகின்றது. சொற்களை உருபன்களாகப் பிரிப்பதிலும் அவற்றிற்குப் பொறுத்தமான பொருண்மை மற்றும் இலக்கணத் தகவல்கள் தருவதிலும் கவனம் தேவை. பின்வரும் எடுத்துக்காட்டு இதைக் தெளப்வுப்படுத்தும்.

தமிழ்சொல் : பணத்திற்காகத்தானே

உருபனியல் பிரிப்பு : பணம் + க்கு + ஆக + தான் + ஏ

வேர் சொல் : (பணம்)

நான்காம் வேற்றுமை உருபு (க்கு)

கொடை வேற்றுமை உருபு (ஆக)

இச்சொல் ஆங்கிலத்தில் மொழிப்பெயர்க்கப்படும் போது இது ஒரே சொல்லாக மொழி பெயர்க்கப்படாமல் கடிச வாந எயமந அடிஇநல டிஇடல என மொழி மொழிபெயர்க்கப்பட வேண்டும்.

4.14.2.2 சொல் வகைப்பாட்டு அடையாளப்படுத்தல் (Parts of Speech Tagging)

சொல் வகைப்பாட்டு அடையாளப்படுத்தி வாக்கியத்திற்கும், சொற்களுக்கும் சொல்வகைப்பாடு தரவேண்டும். இது வாக்கியத்தின் அமைப்பைப் புரிந்துக்கொள்ளவும் வாக்கியத்தின் அமைப்பை உருவாக்கவும் முக்கியமானதாகும்.

தமிழ் வாக்கியம் : அவன் படி ஏறினான்

He staircase Climbed

மேற்கண்ட வாக்கியம் பின்வருமாறு அடையாளப்படுத்தப்படும்.

அவன் < PRO >

படி < N+PRO >

ஏறினான் < V+PST+3SM >

இதில்,

PRO ⇒ Pronoun 'மாற்றுப் பெயர் '

N ⇒ Noun ' பெயர் '

NOM ⇒ Nominative ' எழுவாய்ப் பெயர் '

V ⇒ Verb ' வினை '

PST ⇒ Past tense ' இறந்த காலம் '

3SM ⇒ 3 Person Singular, Masculine 'படர்க்கை ஆண்பால் '

இந்த வேர்ச்சொல்லான 'படி' 'staircase' என்ற பொருளையோ 'read' என்ற பொருளையோ தரலாம். சோல்வாகைப் பாட்டு அடையாளப்படுத்தி சொல் மற்றும் தகவலைப் பயன்படுத்தி இந்த பொருள் மயக்கத்தைத் தீர்க்கும். இவ்வாறு 'படி' என்பதற்குப் பெயர் சொல்வகைப் பாட்டு அடையாளம் தரப்படும்.

4.14.2.3. தொடரியல் ஆய்வு (Syntax analysis)

தொடரியல் ஆய்வின் போது பகுத்தாப்படுகையில் வாக்கிய அமைப்பு தெரிந்து கொள்ளப்படும். பகுத்தாய்வதால் எச்சத்தொடர் எல்லைகள், முன்னுருபு இணைப்பு அடைகள் என்பன தெரிந்துக்கொள்ளப்பட்டு அவற்றின் பொருத்தமான உறுப்புகளுடன் தொடர்புப் படுத்தப்படும். தமிழ் ஒரு சொல் சுதந்திரமான மொழியாகும். இது தொடரியலில் பகுத்தாய்தலைச் சிக்கலாகும். பெயருடன் இணைந்து வரும் வேற்றுமை உருபுகள் வாக்கியத்தில் இவ்வுறுப்புகளின் பங்களிப்பைக் கண்டு பிடிக்கின்றன. ஆனால் வேற்றுமை உருபு இல்லாது இருப்பு அமைப்பு சார் பொருள் மயக்கத்திற்கு வழிவகுக்கும். எடுத்துக்காட்டாக,

தமிழ் வாக்கியம் : 'அவள் அக்காள் வீட்டிற்குச் சென்றாள்'

' She / he Sister house – dative go+pst 2sf'

என்பதற்கு இரண்டு பகுத்துக் குறிப்புகள் கிடைக்கும் :

(அவள்) NP (அக்காள் வீட்டிற்கு) NP சென்றாள் V

(அவள் அக்காள்) NP (vii TT : RKU) NP சென்றாள் V

4.14.2.4 பொருண்மையியல் ஆய்வு (Semantic Analysis)

பல அர்த்தங்கள் உள்ள சொற்களுக்கு ஒரு குறிப்பிட்ட சூழலுக்குத் தகுந்தவாறு சரியான அர்த்தம் கண்டுபிடிக்கப்பட வேண்டும். இலக்குமொழி நிகரன் மயக்கம் நீக்கப்பட அர்த்தம் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும்.

தமிழ் வாக்கியம் 1 : அவன் படி ஏறினான்

He stair case climb-pst-3sm

தமிழ் வாக்கியம் 2 : அவன் படியால் அளந்தான்

He vessel measure-Pst-3sm

மேற்கூறிய வாக்கியங்களில் படி என்பதற்கு இரண்டு அர்த்தங்கள் இருக்கின்றன.

1. Staircase

2. A vessel used to measure rice and wheat.

முதல் வாக்கியத்தில் ஏறினான் என்ற சொல்லுடன் இணைந்து வருகையைப் பயன்படுத்தி படி என்பதன் அர்த்தம் 'staircase' என்பதுதான் என்றும் இரண்டாம் வாக்கியத்தில் அளந்தான் என்ற சொல்லுடன் சேர்ந்து வருகையால் படி என்பதன் அர்த்தம் 'vessel' என்பதுதான் என்றும் நிர்ணயிக்கப்படுகின்றது.

4.14.2.5 தொடரியல் சார் மாற்றல் (Syntactic Transfer)

மொழிபெயர்ப்பு செல்பாட்டில் மாற்றல் அணுகுமுறை (transfer approach) பயன்படுத்தப்படும் போது தொடரியல் மாற்றப்பகுதி தேவைப்படும். ஊள்ளீடு செய்யப்படும் வாக்கியத்தின் அமைப்பு இலக்குமொழியின் பொருத்தமான அமைப்பிற்குத் தகுந்தவாறு மாற்றப்படுகிறது.

தமிழ் வாக்கியம் : ராமன் புத்தகம் படித்தான்

Raman book read

அமைப்பு: Noun Noun Verb

மேற்கண்ட வாக்கியங்களுக்கு நிகரான வாக்கியங்கள் கீழ்வருவன.

ஆங்கில வாக்கியம் : Raman read a book

அமைப்பு :Noun Verb Noun

இந்த மொழிபெயர்ப்பில் Noun Noun Verb என்ற தமிழ் வாக்கியம் Noun Verb Noun என்று ஆங்கிலத்திற்கு நிகராக மாற்றப்படுகிறது. இந்த எடுத்துக்காட்டு மிக எளிமையானதாகும். ஆனால் இச்செயல்பாடு கலவை வாக்கியங்களுக்கு கிடையில் கடினமானதாகும்.

4.14.2.6 அகராதி பொருத்தம் (Dictionary Mapping)

இருமொழி அகராதியைப் பயன்படுத்தி மூலமொழி வேர் சொற்கள் இலக்குமொழியில் அதற்கு நிகரான சொற்களுடன் பொருத்தம் செய்யப்படும். ஒரு மூலமொழி வேர்ச்சொல்லுக்கு இலக்கு மொழியில் பல சொற்கள் கொண்ட தொடர்கள் நிகரன்களாக வரலாம். சூழல் அடிப்படையில் பொருத்தமான சொல் சூழல் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும். வேர்ச் சொல்லின் இலக்கண பண்புக் கூறுகளின் இலக்கணம் அகராதியில் அடங்கி இருக்கும்.

4.14.2.7 உருபனியல் உருவாக்கம் (Morphdogical Generation)

இங்கு சொல்லின் திரிபு வடிவம் வேர் மற்றும் அதன் இலக்கண பண்புக் கூறுகளால் உருவாக்கப்படும்.

தமிழ் வேர்ச்சொல் : ‘படி’

இலக்கண பண்புக் கூறுகள் : ‘Past tense’+ ‘3rd person, Singular, Mascuine’

சொல் வடிவு : ‘ படித்தான் ’

4.14.3 தமிழ்மொழி ஆய்வுக்குத் தேவையான மூலவளமும் கருவிகளும்

4.14.3.1 கருவிகள் (Tools)

4.14.3.1.1 உருபனியல் ஆய்வி (Morphoogical Analyzer)

95 விழுக்காட்டிற்கு மேல் முழுமையுடன் தொழில் நுட்பத்தின் தற்போதைய நிலையைப் பயன்படுத்தி ஒரு திறமையான உருபனியல் ஆய்வி தேவை. தற்போதைய உருபனியல் ஆய்விகள் மேற்கோள்வாய்வாய்பாட்டு (Paradim approach) அணுகுமுறையையும் முற்றுநிலைத் தானியங்கியையும் (Finite State Automata) பயன்படுத்துகின்றனர்.

4.14.3.1.2 அடையாளப்படுத்தி

வாக்கியங்களை அடையாளப்படுத்த முழுதானியக்க, பகுதிதானியக்க, விதி அடிப்படையிலான அல்லது புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தித் தேவை. இவ்வகையான அடையாளப்படுத்தல் இயந்திரமொழி பெயர்ப்புக்கு மட்டுமல்லாமல் தகவல் பிரித்தெடுப்பு, தகவல் மீட்பு உரை சுருக்கம் மற்றும் பிற பல்வேறுபட்ட பயன்பாடுகளிலும் பயன்படுகிறது.

4.14.3.1.3. இலக்கணப் பகுப்பான் (Parser)

ஒரு பகுத்துக் குறிப்பான் சார்பு இலக்கணம் (dependency Gramman) தொடரமைப்பு இலக்கணம் (Phrase Structure Grammer), கிளை இணைப்பு இலக்கணம் (Tree Adjoining Grammar) போன்ற ஏதாவது இலக்கண வடிவமைப்புகள் அடிப்படையில் வாக்கியங்களுக்குப் பகுப்புக் கிளை தரும். ஒரு இலக்கணப் பகுப்பான் தமிழ்மொழிப் பகுப்பாய்விற்கு மிகத் தேவை. இது விதிகள் மற்றும் பகுத்துக் குறிக்கும் இயந்திரத்தை உருவாக்குதலையும் உள்ளடக்கும். சில இயற்கை மொழி ஆய்வுப் பயன்பாடுகள் முழு இலக்கணப் பகுத்தலை வேண்டாது, அவைகளுக்குப் பகுதி இலக்கணப்பகுப்பு போதுமானது. பெயர்த்தொடர் கண்டுபிடிப்பு, எச்சத்தொடர் கண்டுபிடிப்பு என்பன இந்தப் பயன்பாடுகளுக்கு உதவும்.

4.14.3.1.4 பொருண்மையியல் ஆய்வி (Semantic Analyser)

சொல் அர்த்த மயக்கம் நீக்கும் கருவியை உருவாக்கும் வேலை அதன் தொடக்க நிலையில் தான் இருக்கிறது. இந்த வேலை கூடுதலான கவனக் குவிப்பையும் கடினமான முயற்சிகளையும் வேண்டும். மேலும் இது சொற்கள் இணைந்து வரும் தகவலைச் சேகரிக்க வேண்டி பெரிய அடையாளப்படுத்தப்பட்ட விரிதரவை வேண்டும்.

4.14.3.1.5 உருபனியல் உருவாக்கி (Morphological Generator)

நம்மிடம் உருபனியல் ஆய்வுக்கு நன்றாக அமைக்கப்பட்ட முற்றுநிலை இயந்திரம் இருந்தால் அதை உருபனியல் உருவாக்குதலுக்குச் சிறிது மாற்றமின்ற அல்லது சிறிது மாற்றத்துடன் அதே இயந்திரத்தைப் பயன்படுத்தலாம். இவ்வாறு உருவாக்கப்பட்ட உருபனியல் உருவாக்கிகள் திருப்தியான விளைவுகளைக் காட்டுகின்றன.

4.14.3.2 மூலவளங்கள் (Resources)

4.14.3.2.1 அகராதி (Dictionary)

பல அகராதிகள் தமிழில் இருந்தாலும் அவையெல்லாம் மின்வடிவில் இல்லை. இந்த அகராதிகள் இயந்திரம் படிக்கவியலும் வடிவத்தில் இருந்தால்தான் அவற்றைக் கணினி செயல்பாட்டிற்குப் பயன்படுத்த இயலும்.

4.14.3.2.2 மொழி கடந்த அகராதி (Cross -lingual Dictionary)

இது சாதாரண மொழிகள் கடந்த அகராதியைக் குறிப்பிடவில்லை. இயந்திர மொழிபெயர்ப்புக்கு நமக்கு தேவையான அகராதி மொழி கடந்த வேர் அகராதிகளாகும். தனிப்பட்ட மனிதர்களும் நிறுவனங்களும் அவர்களுடைய பயன்பாட்டிற்காக இருமொழி

அகராதிகளை உருவாக்குகின்றனர். இவை இந்த ஆய்வு களத்தில் சூடுபடும் ஆய்வாரள்களுக்கிடையில் பங்கிடப்படவேண்டும்.

4.14.3.2.3 தரவுத்தொகுதி (Corpus)

ஒரு மொழி தரவுத்தொகுதி அம்மொழியின் நடை, சொல், பயன்பாடு போன்றவற்றின் உருப்படுத்தம் ஆகும். இது மேலும் உருபனியல் பகுப்பாவிகள், பகுத்துக் குறிப்பான்கள் மற்றும் உருவாக்கிகள் என்பனவற்றை மதிப்பீடு செய்யும் பரிசோதனைக் கருவியாகச் செயல்படும். இந்திய மொழிகளின் மைய நிறுவனம் சுமார் 15 மொழிகளுக்கு மூன்று மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதிகளை உருவாக்கியுள்ளது. மேலும் சில தரவுத்தொகுதிகளை அடையாளப்படுத்தி உள்ளது. இருக்கின்ற வசதிகளை வைத்துக்கொண்டு பல்வேறுபட்ட மொழிகளுக்கு விரிதரவைச் சேகரிக்க வேண்டி பொதுவான திட்டம் ஊக்கப்படுத்தப்பட வேண்டும். புத்தகங்கள், நாளிதழ்கள், கால இதழ்கள் போன்றவற்றஇ தவிர இணையதளங்களைப் பயன்படுத்தி தரவுத்தொகுதி சேகரிக்கப்பட்ட வேண்டும்.

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட விரிதரவில் (POS Tagged Corpus) உள்ள எல்லா வாக்கியங்களும் சொல்வகைப்பாட்டிற்கு வேண்டி அடையாளப்படுத்தப்பட்டிருக்க வேண்டும். இந்த அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி, புள்ளியியல் அடிப்படையிலான சொல் வகைப்பாட்டு அடையாளப்படுத்தி உருவாக்குவதற்குத் துணைபுரியும். இம்மூலவளத்தின் மீது பல்வேறு வகையான ஆய்வுகளினால் ஒரு மொழிக்குக் கணினி சார் இலக்கணம் உருவாக்க இயலும்.

இணை தரவுத்தொகுதிகள் (Parellel Corpora)

இணைதரவுத்தொகுதி இருமொழிகளுக்கு மாற்றமைவு விதிகளை (transfer rules) உருவாக்க முக்கியமான மூலவளமாகும். இது புள்ளியியல் அல்லது எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறைகள் உருவாக்கத் துணைபுரியும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதற்கு முன் வரிசையாக்க விதிகள் இணைவிரிதரவிகளின் பத்திகள், வாக்கியங்கள் சொற்கள் இவற்றை வரிசையாக்கம் செய்ய வரிசையாக்கக் கருவிகள் தேவைகள். சுப்ல வெளிப்யீட்டாளர்களிடமும் அரசு நிறுவனங்களிலும் இம்மாதிரியான இணைதரவுத்தொகுதிகள் இருக்கின்றன. இந்த இணைதரவுத்தொகுதிகள்

விளைவாக்கம் உள்ள ஒழுங்குமுறைகளை உருவாக்க வேண்டி ஆராய்ச்சியாளர்களுக்குப் பயன்படுத்தத் தரப்படவேண்டும்.

4.14.3.2.4 மாற்றல் இலக்கணம்

மாற்றல் இலக்கணம் (transfer grammar) இருமொழிகளின் தொடரியல் அமைப்புக்கு இடையில் உள்ள முரண்வேறுபாடுகளை அறிந்து கொள்கின்றது. மாற்றல் அணுகுமுறைகளை உருவாக்க இம்மாதிரியான மாற்றல் தொகுதியை உருவாக்குவது கட்டாயமாகும். மொழிகளைப் பற்றிய ஆழ்ந்த அறிவுள்ள மொழியியலார் இம்மாதிரியான மூல வளத்தை உருவாக்கலாம்.

4.14.3.2.5 இணைவமைதி அகராதி

இணைவமைதி அகராதி (Collocation dictionary) மொழியின் மயக்கமான சொற்களின் முறையான சேர்க்கைகளைக் கொண்டிருக்கும். அம்மாதிரியான அகராதி அர்த்த மயக்க நீக்கிற்கு முக்கியமான பகுதியாகும். தமிழுக்கு அம்மாதிரியான மூலவளத்தைச் சில ஆய்வாளர் குழுமங்கள் செய்து வருகின்றன.

4.14.4. முடிவுரை

இயந்திர மொழிபெயர்ப்பு இன்றைய காலகட்டத்தில் முக்கியமான ஒரு நடைமுறை செயல்பாடாக மாறும் நிலையை நோக்கி வளர்ந்து கொண்டிருக்கிறது. இணைய தளத்தில் ஏராளமான தகவல்கள் தமிழிலும் பிறமொழிகளிலும் இருக்கின்றன. குறிப்பாக ஆங்கிலத்தில் பல முக்கியமான தகவல்கள் இணைய் தளத்தில் வலம் வருகின்றன. இத்தகவல்களை தமிழ் மட்டும் அறிந்த நம்மால் புரிந்து கொள்ள இயலாது. இதற்கு வேண்டி உடனடி மொழிபெயர்ப்பு தேவை. இணைய தளத்தில் இருக்கும் ஆங்கிலத் தகவல்களை உடனடியாக இயந்திரத்தினாலேயே தமிழில் மொழிபெயர்ப்பு செய்து கிடைக்கப் பெற்மானால் அது மிகுந்த வரப்பிரசாதமாகும். அத்தகைய திசை நோக்கித்தான் தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகிறது. தமிழிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து தமிழுக்கும் தமிழிலிருந்து பிற மொழிகளுக்கும் தமிழ்மொழிக்கும் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகின்றன. தமிழ்ப் பல்கலைக்கழகம், அண்ணா பல்கலைக்கழகம், சென்னைப் பல்கலைக்கழகம், AUKBC நிறுவனம் தமிழ் இணைய பல்கலைக்கழகம் என்பன தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் ஈடுபட்டுள்ளன. இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technological Development Indian Languages) என்ற திட்டத்தின் கீழ் இந்திய மைய

அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில்நுட்ப அமைச்சின் (Ministry Communication and Information Technology) கீழ்வரும் தகவல் தொழில்நுட்ப துறையால் பொருளாதார உதவிப்பெற்று பல நிறுவனங்களும் பல்கலைக்கழகங்களும் ஆங்கில மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு இந்திய-இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு என்பனவற்றில் முழுமூச்சுடன் ஈடுபட்டுள்ளன. இம்முயற்சிகள் யாவும் இணையத் தளத்திலிருந்து மொழிபெயர்ப்பை இலக்காகக் கொண்டு அமைகிறது. இந்திய மொழிகளின் மைய நிறுவனம் இந்திய மொழிகளின் பேச்சு மற்றும் எழுத்து விரிதரவுகளை திரட்டும் திட்டத்தை பெற்று செயல்படுத்தி வருகிறது.

முற்றிலும் தானியக்க மொழியியல் பெயர்ப்பு இன்றைய காலகட்டத்தில் இயலாத வகையால் கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பை ஊக்குவிக்கிறது. பொதுவான மொழிக்கு அதாவது பொதுவான மொழி உரைக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவது கடினமாகையால் ஒரு குறிப்பிட்ட பொருண்மைகளை வைத்துக் கொண்டு இயந்திர மொழிபெயர்ப்புச் செய்யும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. ஆங்கில-இந்திய மொழிகள், இந்திய மொழிகள்-இந்திய மொழிகள் மொழிபெயர்ப்பு திட்டத்தில் சுற்றுலா மற்றும் உடல் நலம் (Health) என்பன பொருண்மை களங்களாக ஏற்கப்பட்டு மொழிபெயர்ப்பு செய்யும் முயற்சி மேற்கொள்ளப்பட்டு வருகின்றன.

இயந்திர மொழிபெயர்ப்பு என்பது காலத்தின் கட்டாயமாகும். தமிழுக்கான தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் நாம் ஈடுபட்டே ஆக வேண்டும். இல்லாவிட்டால் இருமொழிய (Bilingual) (ஆங்கிலம், தமிழ்) நிலையிலிருந்து ஒரு மொழிய நிலைக்கு (Mondingual)மாறிவருகின்ற நம் மாணவ சமுதாயத்தைக் காப்பாற்ற இயலாது போய்விடும். இணைய தளத்தில் கிடைக்கும் அறிவு மூலங்களைத் தமிழர்களாய நாம் பெறாது போய் விடுவோம். இதுபோல தமிழில் கிடைக்கும் அரிய தகவல்கள் பிற மொழியினவர்களுக்குப் போய்ச் சேராது. தகவல் பன்மடங்காகப் பெருகிவரும் இக்கால கட்டத்தில் தகவல் பரிமாற்றத்திற்கு இயந்திர மொழிபெயர்ப்பை விடையாகும்.

இயல் 5

தரவுத் தொகுதி உருவாக்கம்

5.1 அறிமுகவுரை

அனுபவவாத ஆய்வுகள் எந்த எழுதப்பட்ட அல்லது பேசப்பட்ட உரைகளையும் பயன்படுத்தி அதன் மேல் செய்யப்படுவதாகும். இம்மாதிரியான தனிப்பட்ட உரைகள் பலவகையான இலக்கிய மற்றும் மொழியியல் ஆய்வுகளுக்கு அடிப்படையாக அமைகின்றது. எடுத்துக்காட்டாக, ஒரு செய்யுளின் அல்லது ஒரு புதினத்தின் நடையியல் ஆய்வு அல்லது ஒரு தொலைக்காட்சிப் பேச்சின் உரையாடல் ஆய்வு என்பனவற்றைக் கூறலாம். ஆனால் அனுபவவாத மொழியியலின் ஒரு வடிவுக்கு அடிப்படையாக அமையும். ஒரு குறிப்பிட்ட பகுதியைப் பரிசோதிக்கும் வழிகளிலிருந்து பல அடிப்படையான வழிகளில் வேறுபடும். கொள்கை அடிப்படையில் ஒரு உரைக்குக் கூடுதலான எந்தச் சேகரிப்பு தொகுதியும் தரவுத் தொகுதி அல்லது தரவுத்தொகுதி எனப்படும். தரவுத்தொகுதி என்று தமிழில் அழைக்கப்படும். corpus என்பது Body என்பதன் இலத்தீன் சொல்லாகும். எனவே தரவுத்தொகுதி என்பதை எந்த ஒரு உரையின் உடல் என்று கூறலாம். ஆனால் தரவுத்தொகுதி என்ற சொல் தற்தால மொழியியல் என்ற சூழலில் பயன்படுத்தப்படும் போது எளிய விளக்க உரையைத் தாண்டி கூடுதல் சிறப்பு அர்த்தங்களைக் கொண்டிருக்கும் போக்கு விரிதரவில் காணப்படும்.

தரவுத்தொகுதியின் இரண்டு முக்கிய பகுதிகள்

1. தகவல் உரைநடையின் உரைகள்
2. கற்பனை உரைநடையின் உரைகள்

தகவல் உரைகள் பின்வருவனவற்றை உள்ளடக்குகின்றது:

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

பத்திரிக்கை : இதழாசிரியர் கருத்து (நிறுவனம் சார்ந்தவை, பதிப்பாசிரியருக்கு தனிநபர் கடிதங்கள்)

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

மதம் (புத்தகங்கள், கால இதழ்கள், சமயக்கட்டுரைகள்)

திறமைகளும், பொது போக்குகளும் (புத்தகங்கள் கால இதழ்கள்)

புகழ் வாய்ந்த மரபுச் செய்திகள் (புத்தகங்கள், கால இதழ்கள்)

கடிதங்கள், வாழ்க்கை வரலாறு, நினைவுகள்

பலதரப்பட்டவை (அரசாங்க ஆவணங்கள், அடித்தள அறிக்கைகள் தொழில் நிறுவன அறிக்கைகள், கல்லூரிப் பெயர்ப்பட்டியல் தொழில் நிறுவனம், வீடு)

கற்றல் மற்றும் அறிவியல் எழுத்துரைகள் (இயற்கை அறிவியல் மருத்துவம், கணினித் தொழில் நுட்பம், பொறியியல்)

சமூக மற்றும் நடத்தை அறிவியல்கள் (அரசியல் அறிவியல், சட்டம் கல்வி)

மனிதவியலிலிருந்து உரைகள்

கற்பனை உரைநடையின் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

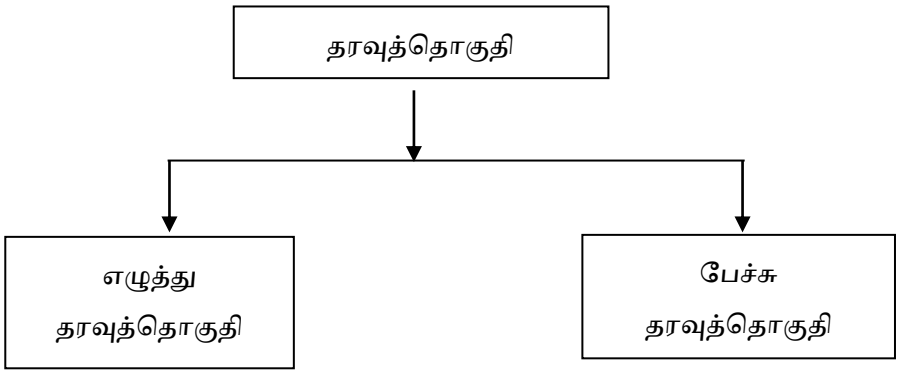
பொதுக் கதைகள் (புதினம், சிறுகதைகள்)

அறிவியல் கதைகள் (புதினங்கள் , சிறுகதைகள்)

துணிக்கரசி செயல் பற்றிய மற்றும் மேற்கத்திய கதைகள் (புதினங்கள் , சிறுகதைகள்)

காதல் கதைகள் (புதினங்கள் , சிறுகதைகள்)

நகைச்சுவை (புதினங்கள் , சிறுகதைகள்)



5.2 ஆங்கில மொழிக்கான தரவுத்தொகுதிகள்/தரவுத்தொகுதிகள்

5.2.1 பிரவுன் தரவுத்தொகுதி

முதல் மின் தரவுத்தொகுதி 1961-ல் பிரவுன் பல்கலைக்கழகத்தில் நெல்சன் பிரான்சிஸ் ஷென்றி குச்சேரா (W.Nelson francies and Henry kuchera) என்பவர்களால் உருவாக்கப்பட்ட

பிரவுண் விரிதரவாகும் (Brown corpus (the standard sample of present day American English). இது 1961-ஆம் ஆண்டு நாள் காட்டி ஆண்டு யுஎஸ்ஏ-ல் USA அச்சடிக்கப்பட்ட ஆங்கில உரை நடையின்ன தொடர்ச்சியான பனுவலின் 1014312 சொற்களைக்கொண்டது. இது 2000ரூ சொற்களைக் கொண்ட 500 மாதிரிகளால் (Sample) பகுக்கப்பட்டள்ளது. ஒவ்வொரு மாதிரியும் ஒரு வாக்கியத்தின் தொடக்கத்தில் தொடங்குகிறது. ஆனால் ஒரு பத்தியின் அல்லது ஒரு பெரிய பகுதியின் தொடக்கத்தில் தொடங்குகின்றது என்பது தேவை இல்லை. ஒவ்வொன்றும் 2000 சொற்களுக்குபின் இறுதியும் முதல் வாக்கியத்தில் முடிகிறது. ஒவ்வொரு மாதிரியும் உரைநடையின் நடைகள் மற்றும் வகைகளின் விரிந்த பரப்பு எல்லையைப் பிரதி நிதித்துவம் 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. முஹ்திரிகள் அவற்றின் பிரதிகளித்துவப் பின்பிற்கேற்ப வேறு எந்த அகவயமாகத் தீர்மானிக்கப்பட்ட சிறப்புக்காகவும் தேர்ந்தெடுக்கப்படவில்லை.

5.2.2 லோப் தரவுத்தொகுதி

லங்காஸ்டர் ஒசலோ பெர்கன் தரவுத்தொகுதி (Lancaster Oslo / Bergen (LOB)) என்பது பிரிட்டிஷ்ஷ் ஆங்கிலத்தின் ஒரு மில்லியன் சொல் சேகரிப்பாகும். இது லங்காஸ்டர் பல்கலைக்கழகத்தைச் சார்ந்த ஜியாபிரே லீச் (Geofferey leech) மற்றும் ஒசலோ பல்கலைக் கழகத்தைச் சார்ந்த ஸ்டிக் ஜோகன்சன் (Slig Johansson) என்போர் பெர்கனில் உள்ள மனிதவியலுக்கான நார்வேஜியன் கணிப்பு மைத்துடன் இணைந்து தொகுக்கப்பட்டதாகும். லோப் யுகே–இல் (UK) 1961-ஆம் வருடத்திற்குள் வெளிப்யிடப்பட்ட மூலப்பெருட்களைக் கொண்டது. இது பிரவுன் விரிதரவைப் போல் உரை மாதிரிகளைக் கொண்டது. லோப் தரவுத்தொகுதியின் திட்ட நெறிமுறை யுஎஸ்ஏ இலும், யுகே-விலும் பயன்படுத்தப்படும் ஆங்கிலத்தின் இரு வகைகளுக்கிடையே உள்ள எதிர்கால ஒப்பீட்டிற்கு வேண்டி பிரவுன் தரவுத்தொகுதிடன் ஒற்றுமைக் கொண்டுள்ளது.

5.2.3 ஆங்கிலத்தில் ஆஸ்டிரேலியன் தரவுத்தொகுதி

ஆங்கிலத்தில் ஆஸ்டிரேலிய தரவுத்தொகுதி (Australian corpus of English (ACE)) என்பது ஆஸ்டிரேலியாவிலுள்ள மாக்குயர் பல்கலைக் கழகத்தில் (Macquaire University) மொழியியல் துறையில் 1986-1990 என்ற கால கட்டத்தில் தொகுக்கப்பட்டது. ஏசியி 1986க்குள் வெளிப்யிடப்பட்டது மூலப் பொருட்களைக் கொண்டிருக்கிறது. ஏசியி பல்வேறுபட்ட மொழியியல் ஆய்வுகளில் உதவுவதற்காகத் திட்டமிடப்பட்ட ஆஸ்டிரேலியாவில் முதலில் தொகுக்கப்பட்ட பலபடித்தான விரிதரவாகும். இது தற்கால ஆஸ்டிரேலிய ஆங்கிலத்தின் திறனான மாதிரியாகும். இது ஆஸ்டிரேலியாவில் கூடுதல் சிறப்புப் பண்புள்ள ஒரு படித்தான தரவுத்தொகுதிகளுடன் ஒப்பீட்டுப் பார்ப்பதற்கு வேண்டி நோக்கிட்டு விரிதரவாகச் செய்யப்படுகிறன்றது. இது பிரவுன் விரிதரவிலும், லோப் விரிதரவிலும் பிரிதி நிதித்துவம் தரப்பட்டுள்ள இனங்களின் சரிசமத்திற்குப் பொருத்துவத நோக்கமாகக் கொண்டது. எனவே இது ஏறக்குறைய ஒவ்வொரு வகைப்பாட்டிலும் 2000 சொல் மாதிரிகளின் நிகரான குழுமத்தை உருவாக்குகிறது. துணை இனங்களையும், விசேஷயப்பரப்புகளையும் பொறுத்துவது தேவை என்பதால் தரவுத்தொகுதி வகைப்பட்டிற்குள்ளும் மாதிரிச் செயல் முறைகள் பெரும்பாலும் திட்டமிடப் பட்டுள்ளது. (வரைமுறையற்றதல்ல)

5.2.4 எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி

எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி (Wellington Written New Zealand English) என்பது 1982 கால கட்டத்தில் வெல்லிங்டன் விக்டோரியா பல்கலைக் கழகத்தில் உருவாக்கப்பட்டது. நியூசிலாந்து ஆங்கில தரவுத்தொகுதியின் நோக்கம், பிரவுன் விரதரவு, லோப் தரவுத்தொகுதி ஏசியி தரவுத்தொகுதி என்பதுடன் நேரடியாக ஒப்பீடு செய்வதை அனுமதிக்கும் எழுத்து நியூசிலாந்து ஆங்கிலத்தின் கணினியாக்கம் செய்யப்பட்ட மாதிரியைத் தருகிறது. முக்கிய உரை வகைப்பாடுகள் எவ்வளவு இயலுமோ அவ்வளவு லோப் தரவுத்தொகுதிடன் பொருந்தும்படி வரிசைப் படுத்தப்பட்டுள்ளது. கற்பனை உரைநடையில் குறிப்பாக கதைகளில் தேவைகருதிச் சில மாற்றங்கள் செய்யப்பட்டுள்ளன.

இந்திய ஆங்கிலத்தில் கோலப்பூர தரவுத்தொகுதி (Kollapur Corpus of India)எஸ்-வி சாஸ்திரியும். அவரைச் சார்ந்தவர்களும் இணைந்து கோலாப்பூரில் உள்ள சிவாஜி பல்கலைக்கழகத்தில் 1988-ஆம் ஆண்டில் வெளிப்யிடப்பட்ட விசயங்களிலிருந்து எடுக்கப்பட்ட இந்திய ஆங்கிலத்தின் ஒப்பீட்டு ஆய்வுகளுக்கு வேண்டிய மூல சாதனங்களாகப் பயன்படுகின்றது. இது இந்திய ஆங்கிலத்தின் விரிந்த வர்ணனைக்குக் கொண்டுச் செல்லும் என்று

எதிர்பார்க்கப்படுகிறது. லோம் மற்றும் பிரவுன் தரவுத்தொகுதிகளை உருவாக்க பயன்படுத்த திட்டத்தைப் பின்பற்றி கோலாப்பூர் தரவுத்தொகுதிக்கு 15 விக்ஷய வகைப்பாடுகளிலிருந்து உரைகள் தேர்ந்தெடுக்கப்பட்டுள்ளன.

5.2.5 ஃப்லோப் தரவுத்தொகுதி

ஃப்லோப் (ஃப்ரெல்பர்க் லோப் தரவுத்தொகுதி (Freiburg lob Corpus for British English (Flob)) மற்றும் லோப் தரவுத்தொகுதிகளுடன் பொருத்தும் ஒரு குழும தரவுத்தொகுதிகளைத் தொகுக்கும் முயற்சியின் விளைவாகும். இது 1990-களின் தொடக்க நிலை மொழியைப் பிரதிநிதித்துவம் செய்யும் என்ற நிலையில் வேறுபடும் இத்திடம் 1990-ஏப்ரலில் ஜெர்மனியில் கிருஷ்டியன் மோரியின் கீழ் தொடங்கப்பட்டது. இது மொழியாலாளர்களுக்குப் பின் வரும் செயல்களில் உதவுகிறது.

- இன்றைய ஆங்கிலத்தில் மொழிசார் மாற்றத்தின் மீதான இன்றைய கருதுகோளாப் பரிசோதிக்க.
- சொல் நிகழ்வெண்களின் முறையான ஒப்பீடுகளின் வழி இலக்கியத்தில் இதுவரை கண்டுப்பிடிக்கப்படாத மாற்றங்களைக் கண்டுபிடிக்க

ஒரு கால இட வழக்குகளின் (British and America) தொடர்ச்சியான மாற்றம் மற்றும் நடையியல் வேறுபாடு இவற்றிற்கு இடையிலான சார்பையும் மெய்யான இருகால வளர்ச்சியையும் ஆயும் முக்கியமான நெறிமுறைச் சிக்கல்களில் ஒன்றை முறையாகக் கையாளுவது. இப்புதிய பிரிட்டிக்ஷ் மற்றும் அமெரிக்க தரவுத்தொகுதிகளின் கூடுதல் பயன் என்னவென்றால் அவை இந்திய, ஆஸ்திரேலிய மற்றும் நியூசிலாந்து தரவுத்தொகுதிகளுடன் (பிந்தைய 1980-களில் உளடள மொழி பயன்பாட்டை பிரதிநிதித்துவம் செய்யும் மாதிரிகள்) ஒப்பிடமூல லோப் மற்றும் பிரவுன் தரவுத்தொகுதிகளைவிடக் கூடுதல் பொருத்தமான தரவு மையங்களைத் தருகின்றன.

5.2.6 பிரிட்டிஷ் தேசிய தரவுத்தொகுதி

பிரிட்டிஷ் தேசிய தரவுத்தொகுதி (British National Corpus (BNC) என்பது மூலங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெறப்பட்டு எழுதப்பட்ட மற்றும் (10%) பல வயதினர்கள், வட்டாரங்கள், சமூக வகுப்புகள் இவற்றிலிருந்து பெறப்பட்ட எழுதப்படாத முறைசாரா உரையாடல்கள் மற்றும் முறையான வணிக மற்றும் அரசாங்கக் கூட்டங்களிலிருந்து வானொலி

மற்றும் தொலைபேசி வரையிலான பரப்பெல்லையைக் கொண்ட சூழல்களின் எல்லா வகைகளிலிருந்தும் சேகரிக்கப்பட்ட பேச்சு மாதிரிகள் என்பனவற்றை உள்ளடக்கும்.

இப்பொதுத்தன்மை பிஎன்சி விரிதரவை அகராதியியல், பொருண்மையியல், உருபனியல், மொழிப் பகுப்பாய்வு, செயற்கை அறிவுநுட்பம், பேச்சு அறிதல் மற்றும் கூட்டிணைப்பு, இலக்கிய ஆய்வுகள் மற்றும் முக்கிய மொழியியலின் எல்லாப் பரப்புகளையும் உள்ளடக்கும் ஆய்வு நோக்கங்களின் விரிந்த வகைக்குப் பயனுள்ளதாகச் செய்கின்றது.

பிஎன்சி பேச்சு உரையாடல் மற்றும் ஒருவர் உரையாடல் இவற்றிலிந்து 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 (code) தரப்பட்டுள்ளது. பிரித்தலும் சொல் வகைப்படுத்தலும் (claws) சொல்வகைப்பாட்டு அடையாளப்படுத்தியால் தானியக்கமாகச் செய்யப்படுகின்றது. பிஎன்சி விரிதரவில் பயன்படுத்தப்படும் வகைப்பாட்டுத் திட்டம் 65 சொல் வகைப்படுகளை வேறுபாடுகளை வேறுபடுத்துகின்றது.

5.2.7 அமெரிக்க தேசிய தரவுத்தொகுதி

அமெரிக்க தேசிய தரவுத்தொகுதி (American National Corpus (ANC)) பிஎன்ச தரவுத்தொகுதி ஒப்பிடத்தக்க அமெரிக்க ஆங்கிலத்தை உட்படுத்திய தரவுத்தொகுதியின் உருவாக்கத்திற்கு உதவியது. இதன் நோக்கம் பிஎன்சி தரவுத்தொகுதிப் பொதுவினங்களுடன் ஒப்பிடத்தக்கக் குறைந்தது 100 மில்லியன் சொற்களைப் பெறுவதாகும். இது மொழித் தொழில் நுட்ப ஆய்வுக்கு உதவுவதுடன் எல்லா மட்டங்களிலும் மொழிப் பகுத்தாய்வுக்கும் கல்விக்கும் பயன்படவேண்டி செழுமையான தேசிய மூலத்தைத் தருகின்றது. எஎன்சி அமெரிக்க ஆங்கில அகராதிகளின் வெளிப்யீட்டார்கள் மற்றும் மொழி ஆய்வில் ஆர்வமுள்ள நிறுவனங்கள் இவற்றின் கூட்டிணைப்பின் (consortium) பங்களிப்பு வழி உருவாக்கப்பட்டது. மொழித் தரவுக் கூட்டிணைப்பும் (linguistics Data consortium LDC) சாதன மற்றும் பொருளாதார ஆதரவும் தந்தது. ஏஎன்சி தரவுத்தொகுதியின் 10 மில்லியன் சொற்களின் முதல் வெளிப்யீட்டில் உள்ளடக்கப்பட்ட உரைகள் முதலில் பெறப்பட்டவையாகும். எனவே, தரவுத்தொகுதி

சரிநிகரானதல்ல பேச்சுப் பகுதி 3224388 சொற்களையும் எழுத்துப்பகுதி 8283828 சொற்களையும் கொண்டுள்ளது.

5.2.8 ஆங்கில வங்கி

ஆங்கில மொழி வங்கி (Bank of English) சொற்கள், இலக்கணம் மற்றும் பயன்பாடு (usage) என்பவனவற்றின் பகுப்பாவிற்கு வேண்டி வடிவமைக்கப்பட்ட தற்கால ஆங்கில மொழியின் மாதிரிகளின் சேகரிப்பாகும். ஆங்கில மொழி வங்கி 1991-ன் (Cobuild (Harper Collins Publishers))மற்றும் யுகேயில் உள்ள பிர்மிங்காம் பல்கலைக்கழகம் இவற்றால் 1991-ல் வெளிப்யிடப்பட்டது. இது புதிய சாதனங்களின் நிரந்தர 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 (brpcjers) துண்டுப்பத்திரிக்கைகள் (leaflets) அறிக்கைகள், கடிதங்கள் மற்றும் பிறவற்றிலிருந்து எடுக்கப்பட்டவைகளாகும். பேச்சுச் சொற்கள் அன்றாட சாதாரண உரையாடல்கள், வானொலி ஒலிபரப்புகள், கூட்டங்கள் நேர்காணல்கள் மற்றும் விவாதங்கள் போன்றவற்றால் பிரதிநிதித்துவம் செய்யப்பட்டுள்ளன.

5.3 இந்திய மொழிகளுக்கான தரவுத் தொகுதிகள்

5.3.1 இந்திய மொழிகளின் MIT தரவுத்தொகுதி

இந்திய மொழிகளுக்கு தரவுத்தொகுதி உருவாக்குவது 1991-ல் தொடங்கப்பட்டது. இது இந்திய அரசின் மின்னுத் துறை எல்லா இந்திய மொழிகளுக்கும் இயந்திரத்தால் படிக்கவியலும் உரைகளின் தரவுத்தொகுதிகளஇ உருவாக்க வேண்டி இந்திய மொழிகளுக்குத் தொழில் நுட்ப வளர்ச்சி (Technological Development for Indian Languages (TDIL)) என்ற திட்டத்தைத் தொடங்கியது. இதே நேரத்தில் மொழி ஆய்வுக்கும் (சொல்வகைப்பாடு அடையாளப்படுத்தி, உரை குறியாக்கி, புள்ளியல் எண்ணி, எழுத்துத் திருத்தி, உருபனியல் பகுப்பாய்வி போன்றவற்றை உருவாக்க) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு எந்திர உதவிசார் மொழிபெயர்ப்புக்கு (machine-aided translation) வேண்டி கருவிகளை வடிவமைத்தல் என்ற செயல்பாடுகளுக்கு மென்பொருள்களை உருவாக்கவும் ஊக்கம் தரப்பட்டது. டெல்கியில் உள்ள இந்தியத தொழில்

நுட்ப நிறுவனம் (Indian institute of technology) இந்திய ஆங்கிலம் , இந்தி மற்றும் பஞ்சாபி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் தமிழ், தெலுங்கு, கன்னடா மற்றும் மலையாள மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. பூனேயில் உள் டெக்கான் கல்லூரி மராத்தி மற்றும் குஜராத்தி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. புபனேஸ்வரிலுள்ள பயன்பாட்டு மொழி அறிவியல்களின் இந்திய நிறுவனம் ஒரியா, வங்காளம், அசாமீஸ் ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. வாரணாசியிலுள்ள சம்பூர்ணநந்தா சமஸ்கிருதப் பல்கலைக்கழகம் சமஸ்கிருத மொழிக்கு தரவுத்தொகுதி உருவாக்கியது. அலிகார் முஸ்லீம் பல்கலைக்கழகம் உருது, சிந்தி மற்றும் காஷ்மீரி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது.

கான்பூரிலுள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் மொழி பகுத்தாய்விற்கும் இயந்திர மொழிபெயர்ப்புக்கும் கருவிகளையும் மென்பொருள்களையும் வடிவமைக்கும் பொறுப்பை ஏற்றது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் எல்லா மொழிகளின் முழு தரவுத்தொகுதி தரவுமையத்தைச் சேகரித்துப் பாதுகாக்கும் பொறுப்பை ஏற்றது. 1995-ல் இத்திட்டத்தின் முடிவுற்றபோது மின்னனு நிறுவனம் இது நீண்டகால ஆய்வையும் பெரிய முதலீட்டையும் வேண்டும் என்று உணர்ந்து திட்டத்தைத் தொடராமல் நிறுத்தியது. இருப்பினும் தற்போது இந்திய அரசின் MIT புதிய வெளிப்ச்சத்திலும் பார்வையிலும் இச்செயல்பாட்டை மீண்டும் நடைமுறைப்படுத்தியுள்ளது.

5.3.2 இந்திய மொழிகளுக்கான தரவுத்தொகுதி

இந்திய மொழிகளுக்கான தரவுத்தொகுதி பற்றி Jayaram மற்றும் Rajashree விரிவாக விளக்கியுள்ளனர். மேலும் இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்தி, பஞ்சாபி, காஷ்மீரி, உருது, கன்னடம், தமிழ், தெலுங்கு, மலையாளம், ஒரியா, அசாமி மற்றும் பெங்காலி போன்ற பன்னிரெண்டு மொழிகளுக்கும் கிட்டத்தட்ட மூன்று மில்லியன் சொற்கள் வருமாறு விரிதரவை உருவாக்கியுள்ளது.

தரவுத்தொகுதியின் நோக்கம்

தரவுத்தொகுதி இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட தரவுத்தொகுதிகள் பல்நோக்கு தரவுத்தொகுதிகள் ஆகும். இந்த ஒரு விரிதரவை உருவாக்க

செலவான தொகை மற்றும் தொகுக்க தேவையான நேரம் என்பன பல்வேறுபட்ட மொழியியல் ஆய்வுகளுக்கு இவற்றை அடிப்படை மூலவளமாக பயன்படுத்த ஊக்கம் அளித்து வருகிறது. ஒருவர் தரவுத்தொகுதி செய்ய எல்லா ஆய்வுகளுக்கும் மிகப் பொருத்தமான முன்மாதிரி என்று எடுத்துக்கொள்ள இயலாது. தரவுத்தொகுதியின் நோக்கம் பின்வருமாறு:

1. வேறுபட்ட உரை வகைகளிலிருந்து சொற்கள் மற்றும் வாக்கியங்களின் மாதிரி.
2. வேறுபட்ட உரை வகைகளில் சொற்களின் நிகழ்வெண்
3. இலக்கண வகைப்பாடுகள் அடிப்படையில் சொற்களின் பட்டியல்
4. வேறுபட்ட சுழல்களில் சொற்களின் தேர்வு
5. தனிநிலை சொற்கள் சேர்ந்து வருகை மற்றும் இலக்கண – தொடரியல் அமைப்பொழுங்கு
6. இலக்கண வகைப்பாடுகள் அடிப்படையில் சொற்களின் பட்டியல்
7. பயன்பாட்டடின் இயற்கையான எடுத்துக்காட்டுகளின் தேர்வுக்குக் கச்சாப்பொருள்.

5.3.3 EMILLE – தரவுத்தொகுதி

EMILLE [Enabling Minority Language Engineering] தரவுத்தொகுதிகள், Emille திட்டம், யு.கே.யில் உள்ள லங்காஸ்டர் பல்கலைக்கழகம், மைசூரில் உள்ள இந்திய மைய நிறுவனம் இவற்றின் ஒத்துழைப்பால் உருவாக்கப்பட்டது. Emille ஐரோப்பிய மொழி மூலவளங்கள் சங்கத்தால் [European Language Resources Association] வழங்கப்பட்டது. இந்த தரவுத்தொகுதி மூன்று பகுதிகளைக் கொண்டது.

1. ஒரு மொழி தரவுத்தொகுதி [Monologual Corpus]
2. இணை தரவுத்தொகுதி [Parallel Corpus]
3. அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி [Annotated Corpus]

ஆசாமி, பெங்காளி, குஜராத்தி, இந்தி, கன்னடம், காஷ்மிரி, மலையாளம், மராத்தி, ஒரியா, பஞ்சாபி, சிங்களம், தமிழ், தெலுங்கு, உருது ஆகிய பதினாங்கு தெற்காசிய மொழிகளின் எழுத்து மற்றும் பேச்சுத் தரவை உள்ளடக்கிய ஒரு மொழி தரவுத்தொகுதிகள் உள்ளன. நுஅடைடந ஒரு மொழி தரவுத்தொகுதி கிட்டத்தட்ட 92,799,000 சொற்களை உள்ளடக்கியது. (இது பெங்காளி, குஜராத்தி, இந்தி, பஞ்சாபி, உருது என்ற மொழிகளின் எழுத்துப்பெயர்க்கப்பட்ட பேச்சுத்தரவின்

2,627,000 சொற்களை உள்ளடக்கியுள்ளது) இணை தரவுத்தொகுதிகள் ஆங்கில உரையின் 2,00,000 சொற்களையும், இந்தி, பெங்காளி, பஞ்சாபி, குஜராத்தி, உருது இவற்றின் மொழிபெயர்ப்புகளையும் கொண்டது. அடையாளப்படுத்தப்பட்ட பகுதி உருது மொழியின் சொல்வகைப்பாட்டு அடையாளப்படுத்தப்பட்ட ஒரு மொழிய மற்றும் இணை தரவுத்தொகுதிகளை உள்ளடக்கியது. இத்துடன் வரையறை பயன்பாட்டின் இயல்பைக்காட்ட இருபது இந்தி தரவுத்தொகுதிகளின் கோப்புகளைக் கொண்டுள்ளது. இந்த விரிதரவானது CES Compliant SGML- ஐப் பயன்படுத்திக் குறிக்கப்பட்டுள்ளது. மேலும் Unicode- ஐப் பயன்படுத்தி குறியாக்கம் (Encode) செய்யப்பட்டுள்ளது.

5.3.4 EMILLE மற்றும் CIIL ஒரு மொழி எழுத்து தரவுத்தொகுதி

எழுத்து தரவுத்தொகுதி டெல்லியில் உள்ள இந்திய தொழில் நுட்ப நிறுவனம் (CIIT), புவனேசுவரத்தின உள்ள பயன்பாட்டு மொழி அறிவியல் நிறுவனம் (ILAS), அலிகாரில் உள்ள அலிகார் முஸ்லீம் பல்கலைக்கழகம் இவற்றுடன் ஒருங்கிணைந்து இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்டது CIIL தரவுத்தொகுதி ஆகும்.

இந்த தரவுத்தொகுதி தொடக்கத்தில் ISCII உரையாக குறியாக்கம் செய்யப்பட்டுள்ளது. இப்பொழுது CES – Compliant SGML குறியீட்டுடன் Unicode – CIIL ஒரு மொழிய தரவுத்தொகுதி 93,530,000 சொற்களைக் கொண்டுள்ளது.

5.3.5 தமிழுக்கான தரவுத்தொகுதி தயாரித்தல்

தமிழுக்கான தரவுத்தொகுதி தயாரித்தலில் முன்னோடியாக உள்ள நிறுவனம் மைசூரில் உள்ள இந்திய மொழிகளின் மைய நிறுவனம் ஆகும். பல பிற நிறுவனங்களும் தாங்கள் எடுத்துக் கொண்ட இயற்கை மொழியாய்வு திட்டங்களுக்கு ஏற்றவாறு தரவுத்தொகுதிகளைச் சேகரித்துப் பயன்படுத்துகின்றன. மின் வடிவில் உள்ள புத்தகங்களையும் கால இதழ்களையும் நாள் இதழ்களையும் தேவைக்கேற்றவாறு சேகரித்து விரிதரவாகப் பயன்படுத்த இயலும். இந்திய மொழிகளின் மைய நிறுவனம் இந்திய மொழிகளுக்காக தரவுத்தொகுதி தயாரிக்கும் முயற்சியில் தமிழுக்கு விரிதரவைத் தயாரித்து தந்துள்ளது.

தமிழுக்கான விரிதரவைப் பொறுத்த வரையில் உரைகளானது தினகரன் பத்திரிக்கையின் இணையத்தளத்திலிருந்து எடுக்கப்பட்டுள்ளது. இந்த தரவுத்தொகுதி இந்திய மொழியின் மைய

நிறுவனத்தால் உருவாக்கப்பட்ட தொடக்கால விரிதரவையும் உட்படுத்தும். தமிழ் எழுத்து விரிதரவில் எடுக்கப்பட்ட சொற்களின் பட்டியல் கீழே கொடுக்கப்பட்டுள்ளது.

1.சினிமா (cinema) திரைப்படம் மற்றும் அது தொடர்பாக செய்திகள் (கிட்டத்தட்ட 1,050,000 சொற்கள்)

2.செய்திகள் (News) கதைச் செய்திகள் (கிட்டத்தட்ட 8,610,000 சொற்கள்)

3.மற்றவை (Others) இணைய தளத்திலிருந்து எடுக்கப்பட்ட பிற கதைகள் (கிட்டத்தட்ட 730,000 சொற்கள்)

4.அரசியல் (Political) அரசியல் செய்தி அறிக்கை (கிட்டத்தட்ட 5,050,000 சொற்கள்)

5.விளையாட்டுகள் (Sports) விளையாட்டுச் செய்தி அறிக்கை (கிட்டத்தட்ட 1,170,000 சொற்கள்)

5.3.6.பிற நிறுவனங்களின் தமிழ் தரவுத்தொகுதிகள்

இயற்கை மொழி ஆய்வில் குறிப்பாக இயந்திர மொழிபெயர்ப்பில் சூடுப்பட்டுள்ள நிறுனங்கள் தமிழுக்கென்று தரவுத்தொகுதிகள் தயாரித்து அவற்றை சொல் வகைப்பாடுகளுக்கு அடையாளப்படுத்தி பயன்படுத்தி வருகின்றன. குறிப்பாகத் தமிழ்ப்பல்கலைக்கழகம், அண்ணாமலைப் பல்கலைக்கழகம், Aukbc ஆராய்ச்சி நிறுவனம், அமிர்தா தொழில் நுட்ப நிறுவனம் என்பன இயந்திர மொழிபெயர்ப்பு ஆய்வில் சூடுபட்டு அதற்கேற்ற தரவுத்தொகுதிகளை உருவாக்கி பயன்படுத்துகின்றன.

5.4 உரை தரவுத்தொகுதி ஆய்வு

பல மொழிகளில் பெரிய மின் தரவுத்தொகுதிகள் சேகரிக்கப்பட்டபின் அவற்றை ஆய்வு உத்திகளுக்கு உள்ளாக்கும் தேவைவரும். விரிதரவிலிருந்து மொழித்தரவை அணுகுவதற்கும் தேவையான தகவல்களை மீளப்பெறுவதற்கும் மக்கள் ஒழுங்குமுறைகளும் உபாயங்களும் உருவாக்கியுள்ளனர். இவ்வாய்வுக் கருவிகள் பயனுள்ளதாக அமைந்துள்ளன. பலவிதமான ஆய்வு உத்திகள் இருக்கின்றன. எடுத்துக்காட்டாக.

1. புள்ளியல் ஆய்வி (Statistical analyzer)

2. தொடரடைவு ஆய்வி (Concordencer)

3. சொல் சேர்ந்து வருகை ஆய்வி (Lexical collocator)

4. முக்கியச்சொல் காணும் ஆய்வி (Key word finder)

5. அகராதிச்சொல் ஆய்வி (lemmantizer)

6.உருபனியல் ஆய்வியும் உருவாக்கியும் (Marphological analyzer and gernerator)

7.சொல் ஆய்வி (word processer)

8. சொல் வகைப்பாடு அடையாளப்படுத்தி (part of speech tagger)

9. தரவுத்தொகுதி குறிப்புரைப்பான் (corpus annotator)

10. பகுத்துக் குறிப்பான் (Parser)

ஆங்கிலம், பிரஞ்சு, ஜெர்மன் போன்ற மொழிகளுக்கு விரிதரவை ஆய பல மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஆனால் இந்தி, தமிழ் போன்ற இந்திய மொழிகளுக்கு அத்தகைய மென்பொருள்கள் இனிமே தான் உருவாக்கப்பட வேண்டும். இதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

5.4.1 நிகழ்வெண் ஆய்வு

மொழியில் நீண்ட காலமாகவே புள்ளியியலுடனும் கணக்கியலுடனும் தொடர்புடையது. மொழியியல், நடை அளவியல் (stylometrics) போன்ற இயற்கைமொழி விரிதரவிலிருந்து பெறப்படும் வேறுப்பட்ட புள்ளியியல் மற்றும் அளவில் தகவல்களை வேண்டுகிறது. மொழியின் அட்டவணைப் படுத்தப்படும். உரையிலுள்ள பல முக்கியமான மொழி அமைப்பொழுங்குகளை அணுக உதவுவதால் இது மிக முக்கியமானதாகும். இது உள்ளுணர்வால் அறிய இயலாத தகவல்களைக் தருகின்றது.

5.4.2 சொல் வருகை ஆய்வி

சொற்களின் சேர்ந்து வருகை உரைகளில் சொற்களின் பங்களிப்பையும் இடத்தையும் புரிந்துக்கொள்ளப் பெரிதும் உதவும் இவ்வாய்வு சொற்களில் எந்த இணைகளுக்கிடையில் கூடுதல் சேர்ந்து வருகை உறவு இருக்கின்றது என்பதை அறிய இயலும். மேலும் இரண்டு சொற்கள் சேர்ந்து வரும் நிகழ்வுத் தகைமையைக் தற்செயலாக விளைவின் அடிப்படையிலான நிகழ்வு தகைமையுடன் ஒப்பிடும் ஒவ்வொரு சொல் இணைக்கும் ஒரு மதிப்பெண் தரப்படும் கூடுதல் மதிப்பெண் சேர்ந்து வருகையைக்காட்டும். இது அகராதியிலும் தொழில் நுட்ப மொழி

பெயர்ப்பிலும் பயன்படுத்த விரிதரவிலிருந்து பல்சொல் அலகுகளைப் பிரித்தெடுக்க உதவும். இயந்திர மொழிபெயர்ப்புக்கும் பொருண்மை வேறுபாட்டிற்கும் இத்தகைய தகவல்களை கைக்கொடுக்கும் இது அர்த்த வேறுபாடுகளை அடையாளம் காண ஒற்றுமையான சொற்களைக் குழும உதவுகிறது.

எ.கா.

மாடிப் - படி Step

வாசற் - படி Step

புத்தகம் படி Read

சொன்னப்படி Manner

ஐந்து லிட்டர் Measure

ஒரே பொருண்மையைக் கொண்ட சொல்லுகளுக்கு இடையே உள்ள பயன்பாட்டு வேறுபாடுகளை வேறுபடுத்த உதவும்.

5.4.3 சூழலில் முக்கியச் சொல்

தரவுத்தொகுதி ஆய்வில் “KWIC” (Key Word in Context (KWIC))’’ என்ற சொல் அடிக்கடி பயன்படுத்தப்படுகிறது. இது குறிப்பிட்ட சொற்களின் நேர்வுகளை அறிய உதவுகின்றது. ஆய்வுக்கு உட்படும் சொல் ஒவ்வொரு வரியின் மத்தியில் கூடுதலான இடை வெளிப்புடன் தோன்றும் வேறுபட்ட நோக்கங்களுக்கு வேண்டி சூழலின் நீட்சி குறிப்பிடப்பட்டிருக்கும். இது மையத்திலுள்ள சொல்லின் இரண்டு பக்கங்களிலும் இரண்டு மூன்று அல்லது நான்கு சொற்களின் சூழலைக் காட்டும். இதை N-Gram (bi-gram, trigram, tetra-gram set) என்பர் இவ்வமைப்பொழுங்கு ஒருவரின் தேவை அடிப்படையில் மாறும். சொற்கள், தொடர்கள், எச்சத் தொடர்கள் இவற்றின் ஆய்வின் போது கூடுதலான சூழல் அவற்றைப்புரிந்துக்கொள்ளத் தேவைப்படும். மறஇஉ என்பதை ஒரு வரையாக எண்ணுவது நல்லது. சொற்களின் நிகழ்வெண்ணை மையச்சொல்லின் சூழலில் பரிசோதிக்கவேண்டும். எல்லாத் தகவல்களுகம் எப்போதும் தேவைப்படுவதில்லை. ஆனால் நாம் தேவைப்படும் போது தகவலைப் பயன்படுத்துகின்றோம்.

KWIC சொல்லால் பொருந்தலை அணுகியபின் நாம் மொழியியல் வர்ணணையில் பல்வேறுபட்ட நோக்கங்களை உருவாக்கலாம் மற்றும் இந்நோக்கங்களை நிறைவேற்ற வழிமுறைகளை வகுக்கலாம். சூழலில் முக்கியத்துவம், சேர்ந்து வரும் சொற்களின் பங்களிப்பு சூழல்களில் சொற்களின் உண்மை நடத்தை, நேர்வின் உண்மைச் சூழ்நிலை மற்றும் சூழல் அடிப்படையிலான கட்டுப்பாடுகள் இவற்றைப் புரிந்துகொள்ள KWIC உதவும்.

குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் 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) பகுத்துக் குறித்தலின் போது உறுப்புகளின் செயல்பாட்டு நடத்தையை நேரிட தகவலைத் தருகிறது.

இது பயன்பாட்டில் வழக்கமற்ற ஆனால் தற்சுட்டு வடிவுகளுடன் தனிப்பட்ட உறவு கொண்ட வினைவடிவுகளின் எடுத்துக்காட்டாக (Amuse onself, please one self, lend on self remind one self)வகைமுறையைத் தீர்மானிக்கின்றது. இம்மாதிரியான அமைப்பொழுங்கின் அறிவு, மொழி கற்பவர்களுக்கு இடைப்பட்ட நிலையிலிருந்து உயர்ந்த நிலைக்குச் செல்ல குழுமுதலைப் பயன்படுத்தி நாம் தமிழில் முற்றும் வினைகள், பெரும்பாலான எச்ச வினைகள் கூடுதலாகத் தொடரும் என்றும் பெயர்கள் ஒட்டுக்களாலும், பின்னுருபுகளாலும் தொடரப்படும் என்றும் அறிந்து கொள்ளலாம். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல்கள் சொற்களை வினைக் குழுக்களாகவும் பெயர்க் குழுக்களாகவும் ஆய உதவும். குறிப்பிட்ட இடம் சாரந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல் சொற்களின் குறிப்பிட்ட இடம் சார்ந்த சேர்க்கையின் காரணமாக வரும் சொல்மயக்கத்தைத் தீர்க்கின்றது. பொருண்மையின் நுண்மையான வேறுபாடுகள் சூழல்களின் சொற்களின் வருகை முறையுடன் கூடிய உறுப்புகளுக்கிடையே உள்ள அக உறுப்புக்காளல் பெரும்பாலும் தரப்படுகின்றது. கூட்டுப் பெயர்களுக்கும், கூட்டு வினைகளுக்கும் சொற்களின் குறிப்பிட்ட சேர்க்கையால் உணர்த்தப்படும் பொருண்மைத் தனிப்பட்ட சொற்களின் பொருண்மையிலிருந்து பெற இயலாது.

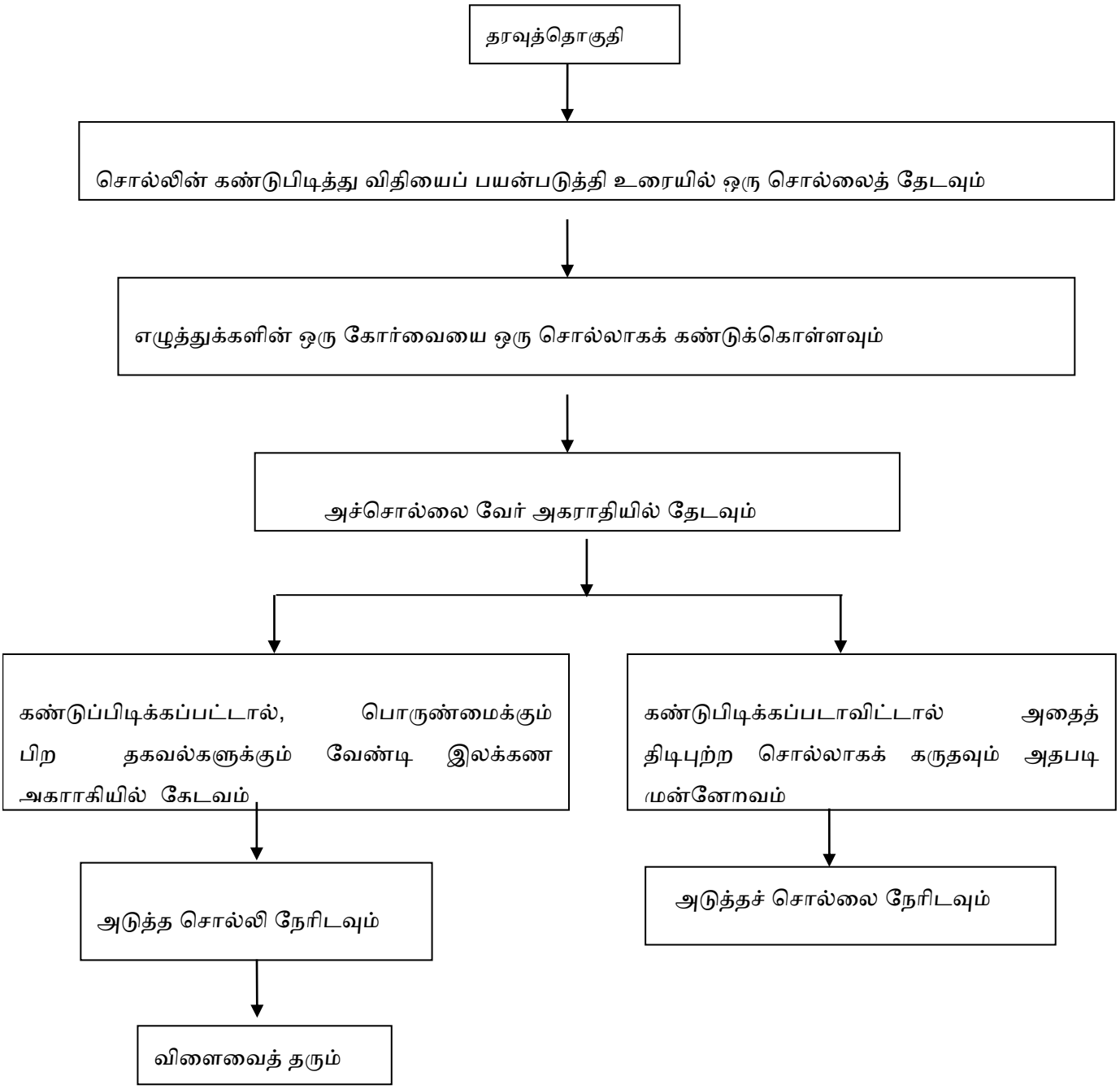
3.4.4 சொல் பகுப்பாய்வு

சொல் பகுப்பாய்வு (Word Processing) விரிதரவில் பயன்படுத்தப்படும் சொற்களைத் தானியக்கமாக ஆய்வதை உள்ளடக்கும். இதன் முக்கிய நோக்கம், ஒரு உரையின் பகுதியிலிருந்து ஒரு சொல்லைக் கண்டுபிடித்து அதை அதன் பயன்பாட்டின் சூழலிருந்து பிரித்து, அதன் உருபொலியனியல் அமைப்பை ஆய்ந்து, அதன் மூலப் பொருண்மையைப் பெற்று, உரையில் அதன் தொடரியல் பங்களிப்பை விளக்குவது அவசியம்.

சொல் பகுப்பாய்விலிருந்து கிடைக்கும் தகவல் சொல் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கும் (Word Sense Disambiguation (WSD))அகராதி உருவாக்குவதற்கும் பகுத்துக் குறிப்பதற்கும் மொழி கற்பத்றகும் மற்றும் இது போன்ற பலற்றிற்கும் பயனுள்ளதாகும். ஆங்கிலம் போன்ற மொழிகளுக்குப் பல சொல் பகுப்பாய்வுகள் உருவாக்கப்பட்டுள்ளன. தமிழ் போன்ற இந்திய மொழிகளுக்கும் இத்ததகைய பகுப்பாய்வு உருவாக்கப்பட்டு பயன்படுத்ததப்பட்டு வருகின்றன.

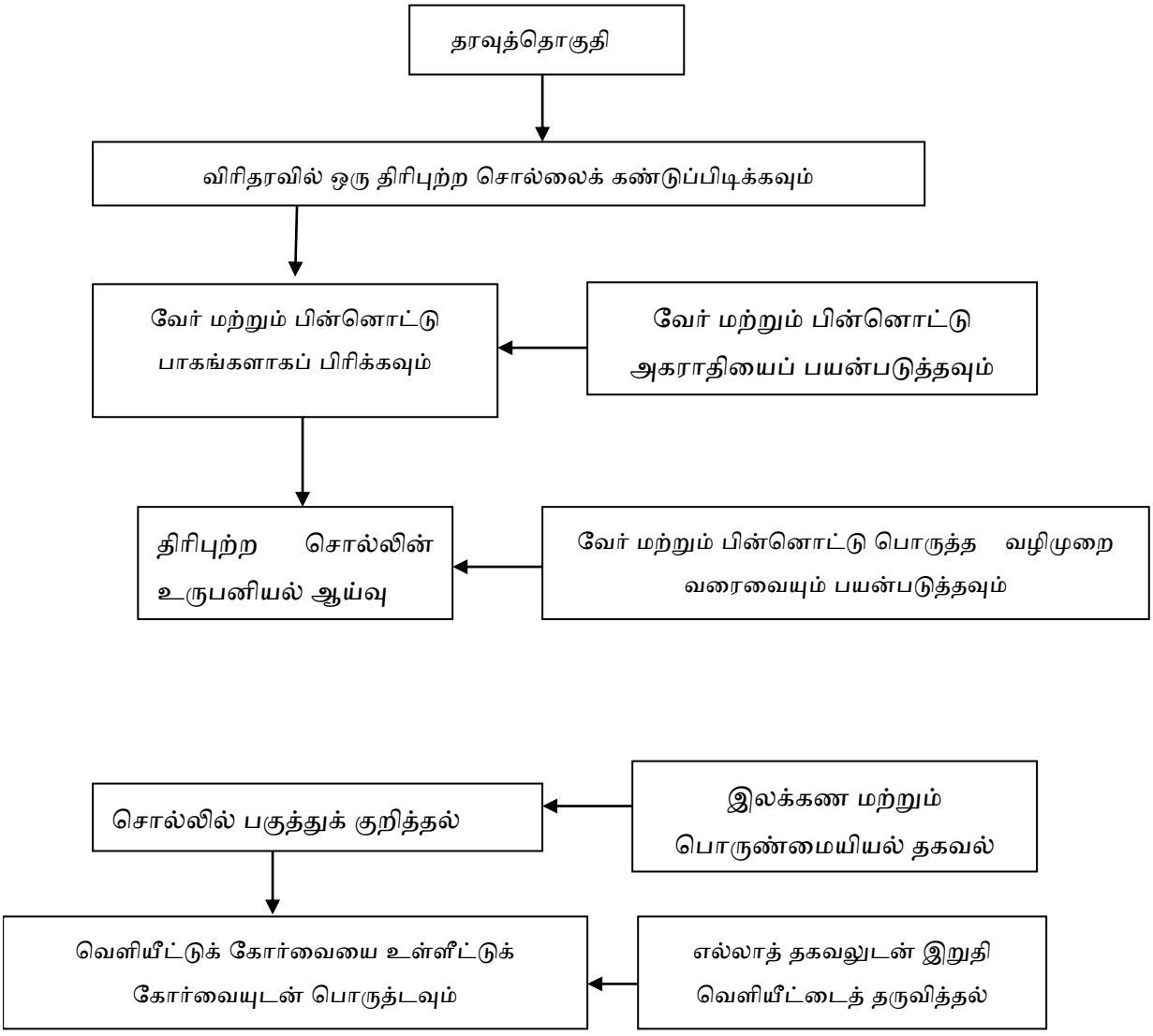
5.4.5 திரிபுறாத சொற்களின் ஆய்வு

திரிபுறாத சொற்களின் ஆய்வு (Processing Non-Inflectecd Forms) பின்வரும் படம் சொற்களின் பகுப்பாய்வை விளக்கும்.


தரவுத்தொகுதி
சொல்லின் கண்டுபிடித்து விதியைப் பயன்படுத்தி உரையில் ஒரு சொல்லைத் தேடவும்
எழுத்துக்களின் ஒரு கோர்வையை ஒரு சொல்லாகக் கண்டுக்கொள்ளவும்
அச்சொல்லை வேர் அகராதியில் தேடவும்
கண்டுப்பிடிக்கப்பட்டால், பொருண்மைக்கும் பிற தகவல்களுக்கும் வேண்டி இலக்கண அகாகியில் கேடவம்
கண்டுபிடிக்கப்படாவிட்டால் அதைத் திடிபுற்ற சொல்லாகக் கருதவும் அதபடி மன்னோவம்
அடுத்த சொல்லி நேரிடவும்
அடுத்தச் சொல்லை நேரிடவும்
விளைவைத் தரும்

5.4.6. திரிபுற்ற சொற்களைப் பகுத்தாய்தல்

பின்வரும்படம் திரிபுற்ற சொற்களின் (Processing Inflecled words) பகுப்பாய்வை விளக்கும்.



5.4.7 இரட்டைச் சொற்களைப் பகுத்தாய்தல்

இரட்டைச் சொற்களின் ஆய்வு (processing dbouble words) உறுப்புகள் அவற்றிற்கு இடையில் ஒரு இடைவெளிப்யால் பிரிக்கப்பட்ட கூட்டுகள் இரட்டுற்ற சொற்கள், வேறுபட்ட சொற்கள் என்பனவற்றை உள்ளடக்கும். இரட்டைசொற்களை விரிதரவில் ஆய்வது மிகவும் சிக்கலான செயல்முறையாகும். பெரும்பாலும் இரட்டைசொற்கள் அகராதியில் தரப்பட்டிருக்கும். அவ்வாறு தரப்படாத சொற்களை காண்பது தான் சிரமமாகும். இதற்கு சேர்ந்து வருகை நிகழ்வெண்கள் இதற்கு பொருந்தும். அடுத்து வருகின்ற பொருண்மை இரட்டைச் சொற்களின் பொருண்மையை உறுதிசெய்ய உயதவுவதால் இரட்டை சொற்களை ஆய்தல் சொல் மட்டத்தில் பொருண்மை மயக்கம் தீர்க்க உதவும்.

5.5 தரவுத்தொகுதி அடையாளப் படுத்துகை

குறிப்பிட்ட பண்புக் குறிப்பிட்டவேண்டிய சொற்களுக்குத் தனிப்பட்ட குறியங்களை ஒட்டுவதை உள்ளடக்கும். மொழிக் குறிப்புரையின் சில வகைகள் குறிப்புரை, குறியங்கள் என்று அறியப் படுவதற்குப் பதிலாக டஇஇபரளைவஇஉ யஇஇடிவயவஇடிஇ அடையாளப்படுத்துதல் என்று அறியப்படுகிறது.

5.5.1 சொல்வகை அடையாளப்படுத்துதல்

சொல் வகைப்பாடு அடையாளப்படுத்தும் (Part Of Speech Tagging) திட்டம் ஒரு சொல்லின் வாக்கியத்தில் அதன் சொல் வகைப்பாட்டுடன் அடையாளப்படுத்தும். இது போன்ற நிலைகளில் செய்யப்படுகிறது. முன்திருத்தல், தானியங்கு அடையாளம் தருகை மற்றும் மனித முயற்சிசார் பின்திருத்தல் என்பவையாகும்.

5.5.2 இலக்கணம் அடையாளப்படுத்தல்

சொல் வகைப்பாடு அடையாளப்படுத்தலில் இரண்டாவது மட்டம் இலக்கண அடையாளப்படுத்தல் ஆகும். இது விரிதரவில் உள்ள ஒவ்வொரு சொல்லுக்கும் இலக்கணம் பொருள் தரும் விரிவான குறிப்புரை ஒழுங்கு முறையாகும். இம்முறையில் ஆங்கிலத்திற்கு சில கருவிகள் உருவாக்கப்பட்டுள்ளது. எம்முறையில் செய்திருதாலும் இறுதியில் மனித முயற்சிசார் திருத்தலுக்கும் மதிப்பீட்டிற்கும் திட்டவரைவு செய்யப்பட்டுள்ளது.

5.5.3 சொற்பொருள் அடையாளப் படுத்தல்

சொற்பொருள் அடையாளப்படுத்தல் (word sense tagging) சொல் அடையாளப்படுத்தப்பட்ட மற்றும் இலக்கணம் அடையாளப்படுத்தப்பட்ட உள்ளீட்டு உரையின் சொற்களை ஏற்க வல்லது. தானியங்கு அடையாப் படுத்தலுக்குப்பின் ஒவ்வொரு சொல்லும் சரியான பொருண்மை வகுப்பாக்கத்தை உறுதி செய்யவேண்டி மனித முயற்சி சார் பின் திருத்தல் செய்யப்படுகிறது.

5.6 மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதி

மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பு இன்றைய காலகட்டத்தில் மிக சிறப்பாகவும் பரவலாகவும் பேசப்படும் மற்றும் பயன்படுத்தப்படும் நெறிமுறையாகும். மொழி புரிந்துகொள்ளுதல், பேச்சு புரிந்துக்கொள்ளுதல், உரை மீட்பு மற்றும் புரிந்துக்கொள்ளுதல் உரைகளிலிருந்து தகவல்களை மீட்டல், ஒலிவழி எழுத்து புரிந்துக்கொள்ளுதல் இயந்திரமொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கும். கணிப்பொறி அடிப்படையிலான மொழியியல் தொழில்நுட்பத்தில் ஆர்வம் கூடி வருகிறது. இருப்பினும் மனித மொழியியல் 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 உருவாக்கியுள்ளது. எல்லா மொழிகளிலும் தரவுத்தொகுதி உருவாக்கப்பட்டு அவற்றை மொழியின் பல நிலைகளிலும் அடையாளப்படுத்தும் முயற்சி நடைப்பெற்று வருகிறது. இத்தகைய தரவுத்தொகுதி மொழிக் கருவிகள் தயாரிப்பதற்கு மிகப்பயனுள்ளதாக அமைகிறது. இந்நெறிமுறை மொழித் தொழில் நுட்பமாக வளர்ந்துள்ளது.

5.6.1 மொழித்தொழில் நுட்பத்தில் தரவுத்தொகுதியின் முக்கியத்துவம்

வேறுபட்ட வகையானன தரவுத்தொகுதி உருவாக்கத்தால் மொழித்தொழில் நுட்ப ஆய்வுப் பயன்பாட்டிலும் தரவுத்தொகுதியின் முக்கியத்துவம் பெருகியுள்ளது. மொழி ஆய்வுக்கான

கருவிகள் மற்றும் ஒழுங்குமுறைகள் உருவாக்குவதில் மொழி தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி உயர்ந்த விளைவைத் தருகிறது. பொதுவாக நாம் தரவுத்தொகுதியின் பயன்பாட்டை இயந்திரத்தால் கட்டுப்படுத்தப்பட்ட கருவிகளைத் திட்டமிடுவதற்கும் தானியங்கு கருவிகளைப் பரிசோதிப்பதற்குப் பயிற்சி தரவும் சிறப்பான மூலவளமாக காண்கிறோம். விரிதரவிலிருந்து பெறப்பட்ட மொழிப் பண்புக்கூறுகளின் நிகழ்வெண்ணிக்கை மொழி கற்போருக்கு நூல்களைத் திட்டமிடவும் ஒளப்வழி எழுத்துக்களைப் புரிந்துகொள்ளும் (OCR) ஒழுங்குமுறைகளை உருவாக்குவதற்கும் தானியங்கு எழுத்துப்பிழைத் திருத்தங்களை உருவாக்குவதற்கும் பயனுள்ளதாக இருக்கிறது. குறிப்புரை செய்யப்பட்ட மற்றும் குறிப்புரை செய்யப்படாத தரவுத்தொகுதிகள் இயந்திரமொழிப்பெயர்ப்பு ஒழுங்குமுறையை திட்டமிடவும் இயந்திரத்தால் படிக்கப்பெறும் அகராதிகளைத் (Machine Readable Dictionaries (MRDS)) உருபனியல் ஆய்விகள், சொல் ஆய்விகள், ஆய்விகள், வாக்கியப் பகுத்துக்குறிப்பான்கள் என்பனவற்றை உருவாக்கவும் பயனுள்ளதாக அமைகிறது. கூப்ரிதரவைப் பயன்படுத்தி உருவாக்கப்பட்ட பல கணினி கருவிகளும் ஒழுங்குமுறைகளும் மொழியைப் பயன்படுத்துபவர்களும், ஆய்வாளர்களுக்கும், எழுத்தாளர்களுக்கும, கல்வியாளர்களுக்கும், ஆசிரியர்களுக்கும், மாணவர்களுக்கும், அறிஞர்களுக்கும், அச்சிட்டு வெளிப்யிடுபவர்களுக்கும், மொழி கற்பவர்களுக்கும் மற்றும் பிறருக்கும் மிகுந்த பயனுள்ளதாக அமைகிறது. பொதுவாக தரவுத்தொகுதியின் முக்கியதுவத்தை மொழித் தொழில்நுட்பத்தின் நான்கு விரிந்த களங்களாகப் பகுக்கலாம் : இது அவற்றின் இயல்பு மற்றும் பயன்பாட்டு நோக்கு அடிப்படைடயில் செய்யப்படும்.

அறிவின் மூலவளமாக தரவுத்தொகுதி

அறிவின் மூலவளமாக தரவுத்தொகுதியின் பரிமாணங்களைப் பின்வருமாறு பட்டியலிடலாம்.

- மொழித் தொழல்நுட்ப கருவிகளையும் ஒழுங்குமுறைகளையும் திட்டமிட மூலவளமாக தரவுத்தொகுதி

- மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளின் மூலவளமாக தரவுத்தொகுதி
- மனித இயந்திர இடைமுக (இடையாக) ஒழுங்குமுறைகளின் மூலவளமாக தரவுத்தொகுதி

அறிவு மூலவளமாக தரவுத்தொகுதி பின்வரும் செயல்பாடுகளுக்குப் பயன்படுகிறது:

- பன்மொழி நூலகங்களை உருவாக்குதல்
- மொழி கற்பவர்களுக்குப் பாடநூல்களைத் திட்டமிடல்
- (அச்சிட்ட மற்றும் மின் வடிவ) ஒரு மொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) இருமொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) பன்மொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) ஒருமொழி சொற்களஞ்சியங்களைப் (பொருட்புல அகராதிகளை) உருவாக்குதல்
- பல்வேறு விதமான (அச்சிட்ட மற்றும் மின் வடிவ) நோக்கீட்டுப் பொருள்கள் உருவாக்குதல்
- இயந்திரத்தால் படிக்கவியலும் அகராதிகளை உருவாக்குதல்
- பன்மொழி சொல் மூலவளங்களை உருவாக்குதல்
- மின் அகராதிகளை உருவாக்குதல்

3.6.2 மொழித் தொழில் நுட்பக் கருவிகளைத் திட்டமிடுதல்

மொழித் தொழில் நுட்பக்கருவிகள் பின்வருவனவற்றை உள்ளடக்கும் :

- சொல்லாய்வு ஒழுங்முறை
- எழுத்துப்பிழைத்திருத்தம் ஒழுங்குமுறை
- உரை நேர்செய்யும் ஒழுங்குமுறை
- உருபனியல் பகுப்பாய்வுஒழுங்குமுறை
- வாக்கியப் பகுத்துக்குறிப்புஒழுங்குமுறை
- நிகழ்வெண் கணக்கிடும் ஒழுங்குமுறை
- சொல் தேடும் இயந்திரம்
- உரை சுருக்கும் ஒழுங்குமுறை
- உரை அடையாளப்படுத்தும் ஒழுங்குமுறை
- தகவல் மீட்கும் ஒழுங்குமுறை

- தொடரடைவு ழுங்குமுறை
- சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை
- சொல் வலை திட்டமிடல்
- பொருண்மை வலை திட்டமிடல்
- சொல் வகைப்பாடு அடையாளப்படுத்தும் ஒழுங்குமுறை
- ஒரிடம்சார் சொற்களைக் குழுமும் ஒழுங்குமுறை

5.7 மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாக தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும் :

1.மொழி மூலவளம்பெறும் ஒழுங்குமுறை

2.இயந்திர மொரிபெயர்ப்புஒழுங்குமுறை

3.பன்மொழிபெயர்ப்புஒழுங்குமுறை

4.மொழி கடந்த தகவல் மீட்வுஒழுங்குமுறை

5.8 மனித-இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கு மூலவளமாக தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்

2. குரலைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்

3.உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள்

4.இணையவலை (இணையத்தளம்) அடிப்படையிலான கற்றல் ஒழுங்குமுறைகள்

5. கேள்விப் - பதில் ஒழுங்குமுறைகள்

6. கணினியின் உதவியுடன் கட்டளைகள்

7 .கணினியின் தவியுடன் மொழிக் கல்வி

8. உரை உருவாக்கம்

5.9 பேச்சுத் தொழில்நுட்பத்தில் தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும் :

1. பேச்சுத் தொழில் நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கம்

2. கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகள்

3. தானியங்கு பேச்சு புரிந்துகொள்ளுதல் செயல்பாடு

4 .தானியங்கு பேச்சு உருவாக்கும் ஒழுங்குமுறைகள்

5. தானியங்கு பகுப்பாய்வு ஒழுங்குமுறைகள்

6. தானியங்கு பேசுபவரை அடையானம் காணும் ஒழுங்குமுறைகள்

7. பேச்சுக் குறைபாடுகள் சரிசெய்தல்

5.10 இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி

இயந்திர மொழிபெயர்ப்பு செயற்கை அறிவு நுட்பத்தின் சிக்கலான ஒரு பகுதியாகும். இது மூல மொழியின் வாக்கியங்களிலிருந்து, இலக்குமொழி வாக்கியங்களை உருவாக்குவதற்கான ஒழுங்குமுறையை உருவாக்கவேண்டும். இலக்கு மொழியின் வெளிப்யீடு உயர்ந்த தரமுடையதாக அமையத் தேவையில்லை. ஒரளவுக்குத் தரம் வாய்ந்ததாய் இருந்தால் போதும். இயந்திர மொழிபெயர்ப்பு மொழிகளுக்கிடையே கருத்துப்பரிமாற்றத்திறகும் தகவல் பரிமாற்றத்திற்கும் மொழிகளைக் கடந்த தகவல் மீட்புக்கும் மொழி கற்றலுக்கும் முக்கியத் தொழில் நுட்பமாகும். பல விதமான பயன்பாட்டின் சாத்தியம் சாதாரணமாக இயந்திர மொழிபெயர்ப்பு மின் வணிகம் தகவல் வட்டாரமாக்கம் பல்மொழி ஆவணப்படுத்தல் மற்றும் தகவல் தொழில் நுட்பம் போன்றவற்றில் மிகப் பயனுள்ள தொழில் நுட்பமாக விளங்குகிறது. நாம் இயந்திர மொழிபெயர்ப்புத் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பைப் பற்றிக் கூறுவதற்கு முன்பின்வருவரும் கேள்விகக்குத் திருப்பதிகரமான விடைதர வேண்டும். மனிதர்கள் மொழி பெயர்ப்பைச் செய்யும்போது ஏன் இயந்திர மொழி பெயர்ப்பை நாம் ஏன் உருவாக்க வேண்டும். இக்கேள்விக்கான பதில் எளிமையானது அல்ல. ஹட்சின் (Hutchin's 1986) பின்வரும் காரணங்களைத் தருகிறார்:

1. பல்வேறுபட்ட அறிவியல் தொழில் நுட்ப ஆவணங்கள் உலகிலுள்ள தொழில் நுட்பவியலாருக்கும் அறிவியலாருக்கும் உடனடியாகத் தரவேண்டுமானால் இயந்திர மொழிபெயர்ப்பு மிகவும் அவசியமாகும்.

2. மனித மொழி பெயர்ப்பாளர்கள் இல்லாத சூழல்களில் இயந்திர மொழிபெயர்ப்பு மிகவும் பயனுள்ளதாக அமையும்.

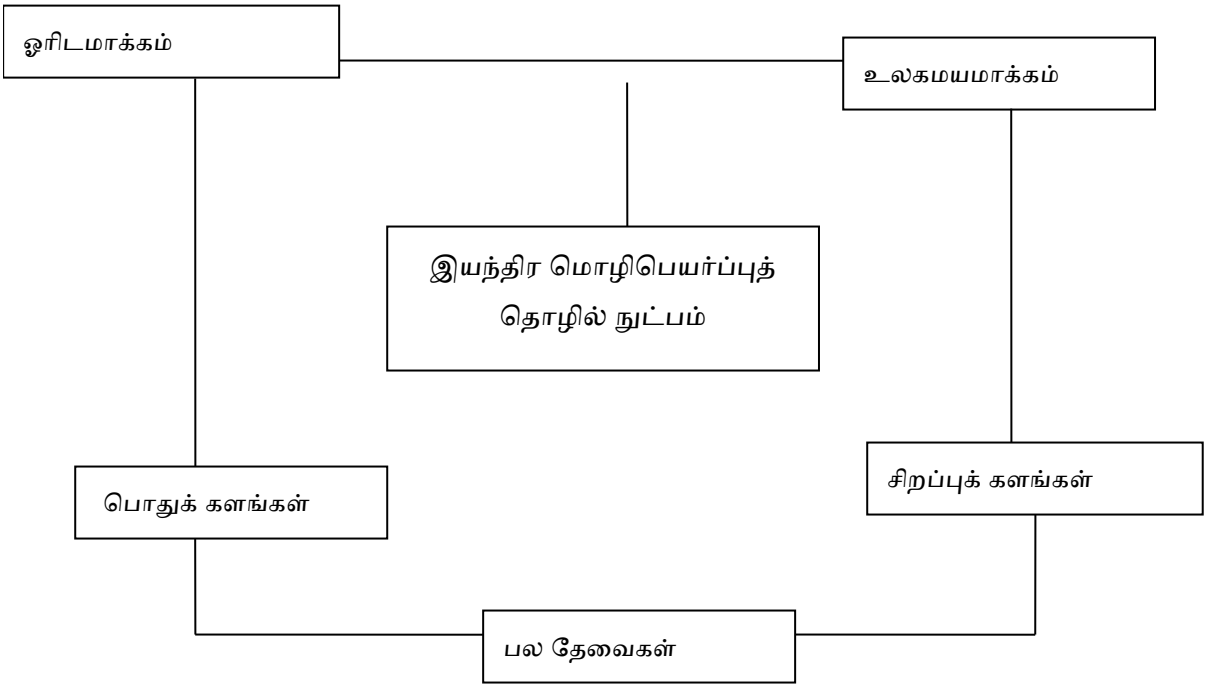
3. இயந்திர மொழிபெயர்ப்பு மொழகளைக் கடந்த ஆவண மாற்றங்கள் மூலம் அக நிலையான கூட்டுறவை மேம்படுத்த இயலும்.

4. இயந்திர மொழிபெயர்ப்பு பல்வேறுபட்ட அறிவியல், தொழில்நுட்பம், விவசாயம் மற்றும் மருத்துவ தகவல்களை ஏழ்மையான மற்றும் முன்னேறுகின்ற நாடுகளுக்கு விரைவான, எளிதான மற்றம் மலிவான பரிமாற்றம் அடிப்படையில் மொழித் தடைகளை நீக்கி அக நிலையான கூட்டுறவையும் மேம்படுத்த இயலும்.

5. மொழிப் பெயர்ப்பு இராணுவ நோக்கங்களுக்கு தொழில்நுட்ப ஆய்வு மற்றும் வணிக நோக்கங்களுக்கு மிகவும் பயனுள்ளது.

1980-களில் இயந்திர ஹுட்சின் இயந்திர மொழிப்பெயர்ப்புக்குச் சாதகமாகக் கூறியபோது இருந்த சூழல் கடந்த ஆண்டுகளில் மிகவும் மாறியுள்ளது. இந்நூற்றாண்டின் தொடக்கத்தில் மக்களின் வாழ்க்கை நடத்தை காரணமாக தாய்மொழி மூலம் தகவலைப் பெறுவது விதியாக மாறிறயுள்ளது. மேலும் படிப்பறிவின் வளர்ச்சி, செய்தி மற்றும் தகவலின் உலகமயமாக்கம், நாடுகளைக் கடந்த பன்மொழிய தரவுத்தொகுதி, இனங்களுக்கிடையே மொழித் தனித்தன்மையின் அவசியம் மற்றும் பிற காரணிகள் கணினி அறிவியலாரையும் மொழியலாரையும் இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உரவாக்க இணைந்து பணியாற்றும் சூழலுக்குத் தள்ளியுள்ளது.

இக்காரணிகள் பங்களிப்பைப் பின்வரும் வரைபடத்தின் மூலம் சுருக்கமாக கூறலாம்.



இங்கு அனுபவவாத தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு முன்மொழியப்படுகின்றது. இது மொழிபெயர்ப்பில் சூடுபடுத்தப்பட்டடுள்ள மொழிகளின் உரைகளிலிருந்து உருவாக்கப்பட்ட மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதிகளின் பகுப்பாய்விலிருந்து பெறப்பட்டத் தகவல்களின் மற்றும் எடுத்துக்காட்டுகளின் அடிப்படையில் அமைந்தது. இங்கு இயந்திர உதவியால் செய்யப்படும் மனித மொழிபெயர்ப்பிற்கு மொழிபெயர்ப்பு தரவுத்தொகுதி பயன்படுத்தப்பட்ட விவாதிக்கப்படுகின்றது. எல்லைப்படுத்தப்பட்ட மற்றும் பொதுக் களங்களில் இவ்வொழுங்குமுறையின் பயன்பாடு அழுந்தக்கூறப்படுகிறன்றது. தகவல் பரிமாற்றத்தின் உலக உயர்வுக்கும் வட்டாரவாக்கத்திற்கும் மொழிபெயர்ப்பு ஒழுங்குமுறையின் தேவை விளக்கப்பட்டு எதிர்கால முன்னேற்றத்திற்கு வேண்டிய வழிமுறைகள் முன்வைக்கப்படுகின்றன. மொழி பெயர்ப்பு வரலாற்றின் மற்றும் தோல்வியல் இன்று கற்றுக்கொண்டவைகளின் அடிப்படையில் இயந்திர மொழிபெயர்ப்பு அமைக்கப்பட்ட ஆலோசனை கூறப்பட்டுள்ளது.

5.10.1 நோக்கம்

கடந்த 50 ஆண்டுகளில் இயந்திர மொழிபெயர்ப்பு தொழில் நுட்பங்கள் மற்றும் பயன்பாடு இவற்றின் அடிப்படையில் கூறத்தக்க அளவு வளர்ந்துள்ளது.இன்றைய இணைய தள கால கட்டத்தில் இணைய தகளத்தில் தகவல்கள் பெரியளவில் பல மொழிகளில் உள்ளன. அவற்றை எல்லா மொழி பேசுவோரும் பகிர்ந்து கொள்ளவேண்டுமானால் அவை ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு உடனடியாக மொழி மாற்றம் செய்யப்படவேண்டும். இதை இயந்திர மொழிபெயர்ப்பின் விரைவான முன்னேற்றத்திற்ன் அடிப்படையில் தான் நிறைவேற்ற இயலும். இது நடைமுறையில் சாத்தியமில்லாத செயல்பாடாக தற்போது தோன்றினாலும் நாம் இதைச் சாத்தியமாகச் செய்கின்ற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதில் ஆவர்ம் காட்டவேண்டும்.

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை ஒரு வடிவிலிருந்து மற்றொருவடிவுக்கு (அதாவது சொல்லிலிருந்து சொல்லையோ கூட்டுச் சொற்களிலிருந்து சுட்டுச் சொற்களையோ தொடரிலிருந்து தொடரையோ வாக்கியத்திலிருந்து வாக்கியத்தையோ) இடப்பெயர்ச்சி செய்யும் அல்லது ஒரு மொழியிலுள்ள உரைகளை மற்றொரு மொழியில் உருப்படுத்தம் செய்யும் எளியமுறை அல்ல. இயந்திர மொழிப்பெயர்ப்பு உருவாக்கப்பட்ட உரை இலக்குமொழியில் இலக்கண அடிப்படையில் சரியான மற்றும் கருத்துரு அடிப்படையில் ஏற்றுக் கொள்ளத்தக்க வகையில் அமைய உறுதி தருவதைக் குறிப்பிடுகின்றது. இயந்தி மொழிப்பெயர்ப்பு அர்த்தம் மற்றும் பொருளடக்க அடிப்படையில் ஒன்றாக இல்லாவிடினும் குறைந்தது மூலமொழிக்கு அண்மையில் இருக்கவேண்டும். மூல மொழியிலுள்ள தகவல் இலக்கு மொழியில் மாற்றப்படும் போது இழப்பு ஏற்படக்கூடாது. மொழிப்பில் மூலமொழியை விடக்கூடுதல் தகவல் சேர்க்கப்படலாகாது. இயந்திர மொழிபெயர்ப்பு உருபனியல், சொல்லியல், தொடரியல் பொருண்மையியல், பயன்வழியியல், கருத்தாடல், புலனறிவு இவற்றைச் சார்ந்திருக்கவேண்டும். இதன் அர்த்தம் என்னவென்றால் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு மனித மொழிபெயர்ப்பாளருக்குள்ள திறனைப் பெற்றிருக்கவேண்டும் என்பதாகும். இதைச் சாதிப்பதற்கு நாம் காத்திருக்கவேண்டும். இருப்பினும் இருமொழி அல்லது இணை மொழி பெயர்ப்பு தரவுத்தொகுதிகள் இதைச் செய்வதற்குச் சாத்தியமான சூழலைத் தருகிறது. இது நம்மை தரவுத்தொகுதிகள் அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு அணுகுமுறைக்குக் கொண்டு செல்கிறது. இதன் காரணமாக இயந்திர மொரிப்பெயர்ப்பு ஒழுங்குமுறையின் திறன் கூடுகிறது.

குறையான தகவலிலன் மூலமாகத் திருத்தப்படாத வடிவில் பயன்படுத்தப்படுவதன் காரணமாகவும் இயந்திர மொழிப்பெயர்ப்பு முற்றிலும் சரியான மொழிப்பெயர்ப்பை உருவாக்குவதை நோக்கமாகக் கொண்டதல்ல எனக் கூறவியலும். சூயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை மொழியில் அடிப்படையிலும் புலனறிவு அடிப்படையிலும் உத்தேச ஒழுங்குமுறையைத்தான் நோக்கமாகக் கொள்கிறது. முயற்சித்தல் மற்றும் பிழை செய்தல் என்ற நீண்ட செயற்பாங்குளின் வழி இலக்கை அடையலாம்.

5.10.2 வரலாற்றிலிருந்து கிடைக்கின்ற பாடம்

1950-களில் இயந்திர மொழி பெயர்ப்பு உலகிலுள்ள பெருமாபாலான எல்லா மொழிகளுக்கும் நேரடியாகவோ மறைமுகமாகவோ பல மூல மொரியிலிருந்து இலக்கு மொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்ய முயற்சிகள் எடுக்கப்பட்டன. இது வரை முழு வெற்றி அடைய இயலவில்லை. அதன் அர்த்தம் என்னவென்றால் மூல மொழிக்கும் இலக்கு மொழிக்கும் மொழி பயன்பாட்டாளர்களால் ஏற்றுக்கொள்ளக்கூடிய மொழிப்பெயர்புகளைத் தானியக்கமாக உருவாக்கும் ஒரு இயந்திர மொழிபெயர்ப்புப் ஒழுங்குமுறை இல்லை. இதன் விளைவாக இயந்திர மொரீபெயர்ப்பு ஆய்வாளர்கள் மரபு சார்ந்த விதி அடிப்படையிலான அணுகுமுறைக்குத் தங்கள் கவனத்தை திருப்பியுள்ளானர்.

5.10.3 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

கடந்த நூற்றாண்டுகளாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கு வேண்டி உருவாக்கப்பட்ட விதிகளின் குழுமங்களால் இயற்கை மொழியின் நெகிழ்வான அமைப்பை உருப்படுத்தம் செய்ய இயலாது. இதன் காரணமாக தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழி பெயர்ப்பு உருவானது. இது பின்வரும் இயந்திர மொழிபெயர்ப்பு வகைகளை உள்ளடக்கும்.

1.எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (example based machine translation)

2.புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (statics based machine translation)

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் விளைவுகளில் ஒன்று இருமொழி உரைகளின் வாக்கியங்களை வரிசைப்படுத்தும் வழிமுறை வரைவுகளின் உருவாக்கமாகும். இவ்வெளிப்ய விளைவு இயந்திர உதவி வழி மனித மொழி பெயர்ப்பின் அடிப்படையில் விக்ஷயங்களில் ஒன்றாக அமைந்தது. ஏனென்றால் இது மொழிபெயர்ப்புக்கு உதவும் பல புதிய கருவிகளுகடகுப் பொருத்தமான அடிப்படையாக அமைந்துள்ளது. மொழிபெயர்ப்பாளர்களால் ஏற்கனவே உருவாக்கப்பட்ட மொழிபெயர்ப்புகளைப் பயன்படுத்தி முழுமையாகவோ பகுதிகாகவோ அவற்றின் அக அமைப்புகளைக் கண்டுபிடிக்க முயலுகிறது. இப்பகுப்பாவு அடிப்படையிலான அணுகுமுறை மொழி பெயர்ப்பாளர்களுக்கு உதவும் கருவிகளைஉருவாக்க உதவியது. இருமொழி இல்லது இணைமொரி பெயர்ப்பு விரிதரவை இயந்திர மொழிபெயர்ப்பில் பயன்படுத்தும் கருத்து புதியது அல்ல. இயந்திர மொழிபெயர்ப்பின் ஆரம்ப காலத்திலேயே இது முயற்சிக்கப்பட்டது. பொதுவாக இணை தரவுத்தொகுதிகள் ஒரு மொழி தரவுத்தொகுதிகளை விட மொழிகளைப் பற்றிய விரிவான தகவலைக் கொண்டிருக்கிறது. தரவுத்தொகுதி ஒழுங்குமுறை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு அணுகுமுறையையும் புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையும் இணைத்து பயன்படுத்துவதால் இது உயர்ந்ததாக இருக்கிறது. விரிததரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மூல மொழியிலும் இலக்கு மொழியிலும் உள்ள தரவுத்தொகுதிகளில் முதன்மையான பகுப்பாயிவிலிருந்து உருவாக்கப்பட்ட வகைகளின் பரப்பெல்லைகளிள் அடித்தளத்தில் அமைந்துள்ளது. பகுப்பாய்வு மொழிபெயர்க்கப்பட்ட விரிதரவில் அடங்கியுள்ள தொடர்கள், மரபுத் தொடர்கள், வாக்கியங்கள், பத்திகள் மற்றும் பிற மொழியியல் பண்புக்கூறுகளின் உருபனியல் பொருண்மையியல் மற்றும் புலனிறவு பொருள்கோள்களை உட்படுத்தும்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் அடிப்படையிலான நெறிமுறை மொழி பெயர்ப்புக்கு முன் வரையறுக்கப்பட்ட தீர்மானங்கள் இல்லை. ஆனால் மனிதரால் மொழி பெயர்க்கப்பட்ட உரைகளில் மிக சாத்தியமான தீர்மானங்கள் காணப்படும் என்ற நம்பிக்கையின் அடிப்படையில் அமைவதாகும். அதாவது மனித மொழிபெயர்ப்பாளர்கள் திறமையின் பெரும்பகுதி மொழி பெயர்க்கப்பட்ட உரைகளில் காணப்படும் மொழி நிகரன்களில் குறியாக்கம் செய்யப்பட்டுள்ளன என்று கூறலாம். எல்லைக்குட்பட்ட களங்களில் இவ்வணுகுமுறையால் பெறப்பட்ட சமீபகால வெற்றி பொதுவான களங்களிலும் இம்மாதிரியான

வெற்றையைப் பெற மொழி சார்ந்த மற்றும் மொழிக்குப் புறம்பான அறிவுதேவை என்று காட்டுகினறது. எல்லாக் களங்களிலும் இவ்வணுகுமுறையின் வெற்றி குறித்து தீர்மானமாக வருவதுரைக்க இயலாது. இருப்பினும் நன்றாகத் திட்டமிடப்பட்டது மொழிப்பெயர்ப்பு விரிதரவிலிருந்து பெறப்பட்ட தகவலிலிருந்து உருவாக்கப்பட்ட நெறிமுறை எல்லைக்குட்பட்ட மற்றும் பொது வளங்களில் வெற்றையைத் தரும் என்று கூறு இயலும்.

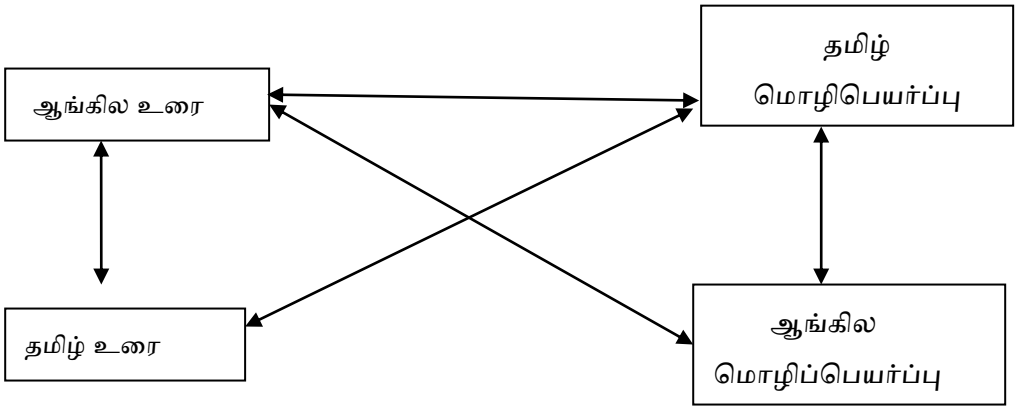
5.10.4 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறையுடன் தொடர்புடைய சிக்கல்கள்

தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு விரிவான தரவுத்தொகுதி பகுப்பாய்வின் மற்றும் ஆய்வின் கூறத்தக்க முன்னேற்றத்தைப் பெற்றுள்ளது. இந்நாள் வரை பல்வேறு வகையிலான மொழிபெயர்ப்பு ஒழுங்குமுறைகள் திட்டமிட உதவும் முக்கியமான உள்ளறிவைத் தரவேண்டி உருவாக்கப்பட்டுள்ளன.. பொதுவாக இம்மொழிபெயர்ப்பு தரவுத்தொகுதி இயற்கையாகப் பெறப்படும் மொழித் தரவுகளின் பெரிய சேகரிப்பை உருப்படுத்தும் செய்கின்றது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மொழிபெயர்ப்பு தரவுத்தொகுதியின் பயன்பாடு கட்டாயமானதாகும் என்று தெளப்வாக்குகின்றது. ஏனென்றால் அவை மொழி மற்றும் மொழிக்குப் புறம்பான எடுத்துக்காட்டுகளின் மற்றும் தகவல்களின் பல்வேறு விதமான வகைகளை நமக்குத் தருகிறது.

5.10.5 மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கம்

மொழிப்பெயர்ப்பு தரவுத்தொகுதி மூல மொழியின் உண்மையான பண்புக் கூறுகளையும் இலக்கு மொழியிலிருந்து பெறப்பட்ட அவற்றின் மொழி பெயர்ப்புகளையும் கொண்டிருக்கும். இவ்தரவுத்தொகுதி மொழிகளைக் கடந்து சொற்களின் மற்றும் தொடர்களின் பொருண்மையையும் செயல்பாட்டையும் கொண்டிருக்கும். இதன் காரணமாக ஒரே கட்டுபாட்டிற்குள் இரு வேறுபட்ட மொழிகள் குறிப்பிட்ட பொருண்மைகளின் மெய்யுருவாக்கத்தை ஒப்பிடுவதற்குப் பொருத்தமான அடிப்படையைத் தருகின்றது. மேலும் அவை மொழிகளைக் கடந்த வேறுபட்ட சொற்களைக் கண்டுபிடிப்பதைச் சாத்தியமாக்குகிறது. இவ்வாறு மொழிப்பெயர்ப்பு தரவுத்தொகுதி மொழிகளைக் கடந்த தரவுப் பகுப்பாய்வுக்குத் தருகின்றது. மொழிபெயர்ப்பு தரவுத்தொகுதியின் உருவாக்கம் ஒரு சிக்கலான வேலையாகும். இது தரவுத்தொகுதியின் உருவாக்கம் மற்றும் கவனமான வழிகாட்டலை வேண்டுகிறது. மொழிபெயர்ப்பு தரவுத்தொகுதி ஒப்பீட்டு இணை தரவுத்தொகுதிகளின் முன்னேற்றத்திற்கான வழிகளை ஒன்றிணைக்க வேண்டி தகுதியுடையதாக

ஆக்க உருவாக்கப்படும் இரு மொழிகளிலிருந்து உரை மாதிரிகள், உரை வகை, பாடப்பொருள், நோக்கம் மற்றும் கலைச்சொல் இவற்றின் அடிப்படையில் இயன்ற வரைக்கும் பொருத்தப்படும் இரு மொழிகளுக்கிடையிலுள்ள மொழிப்பெயர்ப்பு தரவுத்தொகுதியின் அமைப்பு பின்வருமாறு அமையும்.



இம்மாதிரியான தரவுத்தொகுதிக்கு உரை மாதிரிகள் தெரிவு பின்வருமாறு 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.மொழிப்பெயர்ப்பு தரவுத்தொகுதி எந்தக் குறிப்பிட்ட வட்டார மொழிக்கோ மொழி வகைக்கோ உரிய உரை வகைக்கு எல்லைப்படுத்தப்படவில்லை. அவை மொழி பயன்பாட்டின் எல்லாச்

சாத்தியமான களங்கள் மற்றும் துறைகளிலிருந்து பெறப்பட்ட உரை வகைகளின் பரந்த பரப்பெல்லையை உட்படுத்தும்.

4.இரண்டு மொழிகளிலிருந்து எடுக்கப்பட்ட உரைகள் இயன்றவரை ஒப்பிட தக்கவையாக இருக்கும். இவை இனம் (எ.கா.செய்தி), வகை (எ.கா. அரசியல்), பொருளடக்கம் (எ.கா.தேர்தல்) மற்றும் வடிவம் (எ.கா.அறிகை) இவற்றின் அடிப்படையில் பொருத்தம் பார்க்கப்படும். அவைகள் நோக்கம் பயன்படுத்துபவரின் மொழி இவற்றின் அடிப்படையிலும் பொருத்தமுடையதாக இருக்கும்.

5.மொழிபெயர்ப்பு விரிதரவில் உட்படுத்தப்படுள்ள மாதிரிகள் ஒரு இயல்பான தடை நிலையில் (எ.கா. பகுதி, துணைப்பகுதி, அல்லது பத்தி போன்றவை) தொடக்கத்திலிருந்து இறுதிவரை எடுக்கப்பட்ட உரைகளில் விரிவான மற்றும் இயற்கையான பாகங்களைக் கொண்டிருக்கும்.

5.10.6 மொழிபெயர்ப்பு தரவுத் தொகுதியைப் பொருத்தமாக வரிசைப்படுத்தல்

மொழிபெயர்ப்பு தரவுத்தொகுதிகளைப் பொருத்தமான வரிசைப்படுத்துவது என்பது மூலமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு ஆகும். இலக்கு மொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகுடன் பொருந்தவேண்டும் என்பதாகும். இங்கு மொழிபெயர்ப்பு அலகு என்பது சிறிய தொடர்ச்சிகளான சொற்கள், கூட்டுச்சொற்கள், தொடர்கள் மற்றும் வாக்கியங்கள் என்பனவற்றுடன் பெரிய தொடர்ச்சியான பத்திகள், பகுதிகள் என்பனவற்றையும் உள்ளடக்கும். மொழிபெயர்ப்பு அலகுகளின் தெரிவு மொழியில் ஆய்வுக்குத் தெரிந்தெடுக்கப்பட்டுள்ள பார்வையையும், பயன்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகையையும் பொறுத்து அமையும். மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதி மூலத்திலிருந்து (எ.கா.சட்டம் மற்றும் தொழில் நுட்ப தரவுத்தொகுதிகள்) மாறாத உயர்ந்த நிலையை வேண்டினால் இரு தரவுத்தொகுதிகளின் நெருங்கிய வரிசைப் பொருத்தம் வாக்கியங்கள், சொற்கள் என்ற அளவில் விலகி நிற்கும். தரவுத்தொகுதி தழுவலாக இருந்தால் மூலத்தின் நேரடி மொழிபெயர்ப்பாக அமையாமல் பத்திகள், பகுதிகள் என்ற பெரிய அலகுகளை வரிசைப்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகை அடிப்படையில் நேர்படுத்தப்படும். வரிசையாக வருவதும் மனித மொழிபெயர்ப்பாளர்களின் நம்பகத் தன்மையும் மொழிபெயர்ப்பட்ட தரவுத்தொகுதிகளை வரிசைப்பொருத்தம் செய்ய உதவும். இது கூடுதல் தொழில்நுட்ப தரவுத்தொகுதிக்குப் பகுதி உண்மையாகும். முஹ்றாக, இலக்கிய தரவுத்தொகுதிகள், விரிதரவில் பயன்படுத்தப்பட்டு நிகரன்கள் முன்னரே முறையாக்கம்

செய்யப்பட்டிருந்தால் வாக்கிய மட்டத்திற்கு கீழே உள்ள அலகுகளின் வரிசைப் பொருத்தத்திற்கு உட்படும். எந்த நிலையிலும் (பத்தி, வாக்கியம், சொல் போன்ற நிலைகளில்) மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதிகள் இணையான அலகுகள் கொண்ட எளிய சொல் தரவு மையங்களாக் கருதப்படுகிறது. இதன் முக்கிய நோக்கம் இரண்டு மொழிகளுக்கிடையே காணப்படும் அமைப்பு ஒற்றுமைகளைக் காட்டுவதற்கு அல்ல: பயன்வழியில் அடிப்படையில் மூல உரை அலகுகளுக்கு அண்மைப்பட்டதாகக் தோன்றும் இலக்கு உரை அலகுகளைத் தேடுவதாகும். இதைக் செய்வதற்குத் தொடக்கநிலை, இருமொழிய அகராதிகளின் உதவியால் சொற்களின் ஆரம்ப வரிசைப் பொருத்தத்தைச் செய்வதாகும். இம்மாதிரியான குறை வரிசைப் பொருத்தங்கள் வாக்கிய நிலைகளில் திருப்திகரமான விளைவுகளைத் தரும்.

5.10.7 இந்திய மொழிகளின் இன்றைய நிலை

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பில் முக்கியமான தடைகளில் ஒன்று இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கதபகும். ஒருமொழிய விரிதரவில் போதுமான அளவு எழுதப்பட்ட உரைகள்ள இருந்தாலும் பின்வரும் செயல்படுத்துநிலை சிக்கல்களின் காரணமாக இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளை மறுதிட்டவரைவு செய்வதற்கு ஆர்முள்ள முயற்சிகள் எடுக்கப்பட்டவில்லை.

5.10.8 உருவாக்கம்

இந்திய மொழிகளுக்கிடையே வேறுபட்ட வகையான. மொழிப்பெயர்ப்பு தரவுத்தொகுதி 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.) (இயவஇடிஇயட bடிடிம வசரளவ) மற்றும் ளுயாவைலய ஆஉயனநஅல போன்றவற்றால் வெளிப்யிடப்பட்ட பல உரைகள்) ஏற்கனவே உருவாக்கப்பட்டுள்ளது என்றாலும் அவை ஏற்றுக்கொள்ளக்கூடிய மின்விடிவில் இல்லை.

மாற்றம்: வேறுபட்ட மூலம், உரைவகை, மாதிரி மற்றும் வடிவமைப்பு கொண்ட பிற தரவுத்தொகுதிகள் இணை மொழிபெயர்ப்பு தரவுத்தொகுதி மாற்ற வேண்டி ஆயப்படவேண்டும்.

சுத்தமாக்கல்: மொழிபெயர்ப்பு தரவுத்தொகுதி எதிர்கால செயற்பாங்கிற்கும் ஏற்றுக்கொள்ளலுக்கு பயன்பாட்டிற்கும் வேண்டி இயந்திரப் பயன்பாட்டு வடிவமைப்பில் வைத்திருக்கமொழி வல்லுனர்களால் முறையாக தலையிடப்படவும் கட்டுப்படுத்தப்படவும் வேண்டும்.

இணைப்பொருத்தமாக வரிசைப்படுத்தல்: மொழிகளுக்குடையில் பொருத்தும் பகுதிகளை அடையாளங்கான வேண்டி மொழிபெயர்ப்பு தரவுத்தொகுதி இணைப்பொருத்தமாக வரிசைப்படத்தப்படவேண்டும்.

5.10.9 மொழிபெயர்ப்புத் தரவுத்தொகுதிகளில் மொழியின் செயல்பாடுகள்

மொழிப்பெயர்ப்பு தரவுத்தொகுதி தொகுத்தலுக்கும் வரிசைப்படுத்தலுக்கும் பின்னர் மொழியில் பகுப்பாய்வின் வேறுபட்ட நிலைக்கு உள்ளாக்கப்படுவதற்கு முன்னர் பொருத்தமுஹ்ன பல நிலைகளைக் கடக்கும். இது மொழி பெயர்ப்பு நிகரன்கள் முறையாக்கத்தின் வடிவாக்கத்தின் நிலைகளுக்கு அடிப்படையாக அமையும். பொதுவாக மொழியியல் ஆய்வு பின்வருவனவற்றை உள்ளடக்கும் :

- உருபன்களின் வடிவத்தையும் செயல்பாட்டையும் அடையாளம் காண உதவும் சொற்களின் உருபனியல் ஆய்வு
- தொடர்புள்ள தரவுத்தொகுதிகளின் வடிவத்தையும் செயல்பாடுகளையும் அடையாளம் காண உதவும் தொடரியல் ஆய்வு
- தரவுத்தொகுதிகள் பயன்படுத்தப்பட்டுள்ள சொற்களின் புறவடிவங்களுக்கு இடையில் உள்ள இடைமுகத்தை ஆயு உதவும் உருவனியில் - தொடரியல் ஆய்வு. இதன் துல்லியமான மற்றும் திறமையான ஆய்வு பகுப்பாய்வின் தன்மையையும், விரிவையும் அதிகரிக்கும்.
- சொற்கள், தொடர்கள் போன்ற அலகுகளின் பொருண்மையைக் கண்டுபிடிக்கவும் இதில் வரும் பொருள் மயக்கங்களைக் கண்டுபிடிக்கவும் உதவும் பொருண்மையில்.

5.10.10 மொழிபெயர்ப்பு ஆய்வு

இயந்திர மொழிப்பெயர்ப்பு ஆய்வாளருக்கு இடையிலான முக்கியமான வாக்குவாதம் மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளின் ஆய்வில் உட்படுத்தம் கலவைத் தன்மையின் நிலையைப் பற்றியாகும். பொதுவான நம்பிக்கை இயற்கையான உரைகளில் வரும் பல எண்ணிக்கையிலான மொழியியல் நடத்தைகள் ஆயப்பட்டு வெளிப்ப்படையாக உருப்படுத்தம் செய்யப்படும் வரை

உயர்ந்த தன்மையுள்ள இயந்திர மொழிபெயர்ப்பு எய்த இயலாததாகும் என்பதாகும். சொல் மயக்கம் மற்றும் உறுப்பு பொருத்தம் இவற்றின் சிக்கல்களை மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்டு உட்படுத்தப்பட்ட ஒரு மொழியின் அகராதியிலும் இலக்கணத்திலும் சேகரித்து வைக்கப்பட்டுள்ள ஏராளமான அறிவின் உதவியால் தீர்க்க இயலும். இது மொழிபெயர்ப்பு ஆய்வின் கடுமையான செயற்பாங்கின் பயன்படுத்தலை வேண்டும். சுருக்கமாக இயந்திர மொழிபெயர்ப்பில் மொழி ஆய்வு நுட்பத்திற்குப் பல பயன்பாடுகள் இருக்கின்றன. அவை பின்வருமாறு அமையும்:

- மொழிப்பெயர்ப்பு பகுப்பாய்வு முன்னரே இருக்கக் கூடிய மொழி பெயர்ப்புகளை அமைப்புகளை அமைப்பாக்கம் செய்ய உதவும். இதன் மூலம் அவற்றை புதிய மொழி பெயர்ப்புகளின் உருவாக்கத்தில் மறு உபயோகம் செய்ய இயலும்.
- மொழிப்பெயர்ப்பு பகுப்பாய்வு தொழில் நுட்பம் மொழிப்பெயர்ப்பு தவறுகளின் சில வகைகளைக் கண்டுபிடிக்க வேண்டி தொடக்கநிலை மொழிபெயர்ப்புகளில் பயன்படுத்தப்படுகிறது.
- மொழிப்பெயர்ப்பு பகுப்பாய்வு நுட்பம் இணை மொழிச் சொற்களால் ஏற்படும் குறுக்கீட்டுத் தவறுகளிலிருந்து சுதந்திரமாக மொழிபெயர்ப்பு இருந்தால் அதைப் பரிசோதிக்கப் பயன்படுத்தப்படும்.

5.10.11 இருமொழிய அகராதியின் உருவாக்கம்

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் மற்றொரு பயனுள்ள மற்றும் முக்கியப் பகுதி இருமொழி அகராதிகளின் உருவாக்கமாகும். இதன் குறைபாடு இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பின் தடைகளில் ஒன்றாகும். மரபு அகராதியில் இக்குறைபாட்டை நிவர்த்தி செய்ய இயலாது. ஏனென்றால் அவை சொல் துணை வகைப்பாடு, சொல்தெரிவு, சொல்தேர்வுக் கட்டுபாட்டு மற்றும் சொல் அலகுகளின் பயன்பாட்டு களங்கள் இவற்றைப் பற்றிய போதுமான தகவல்களைக் கொண்டிருக்கவில்லை. இருமொழிய அகராதி உருவாக்கத்திற்கு நாம் பல வழிகளில் அடையாளப்படுத்தப்பட்ட மற்றும் பகுத்துக் குறிக்கப்பட்ட தரவுத்தொகுதிகளில் பயன்படுத்தப்பட்டுள்ள பல்வேறு புள்ளியியல் நெறிமுறைகளைப் பயன்படுத்த வேண்டும்.

5.10.12 மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு

மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளில் மொழிப்பெயர்ப்பு நிகரன்களைத் தேடல் மூல மொழியில் ஒரு குறிப்பிட்ட பொருண்மையை அல்லது கருத்துருவை வெளிப்படுத்தும் ஒரு குறிப்பிட்ட வடிவிலிருந்து தொடங்கப்பட வேண்டும். மூல மொழியில் தேடல் செயற்பாங்கின் 2-வது பகுதி தொடங்கும். இது இலக்குமொழியில் ஒரே அர்த்தத்தைக் கொண்டிருக்கிற சொல்லின் அடையாளம் காணலை உட்படுத்தும். இயல்பாகப் பெரும்பாலான மொழிபெயர்ப்பு தரவுத்தொகுதிகள் மொழிபெயர்ப்பு நிகரன்களின் மிகுந்த பரப்பெல்லையை வெளிப்படுத்த இயலும். இவை மாற்று வடிவின் தேர்வை உறுதி செய்யும் காரணிகள், சொற்கள் பயன்பாட்டின் மறு நிகழ்வு அமைப்பொழுங்கின் அடிப்படையில் உறுதி செய்யப்படும்.

5.10.13 கலைச்சொல் தகவல் வங்கியின் உருவாக்கம்

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு தொழில் நுட்ப மற்றும் அறிவியல் கலைச்சொற்களில் பொருத்தமான ஆக்கத்திற்குக் கலைச்சொல் ஆக்கவியலாரின் எண்ணத்தின் அடிப்படையிலான ஆய்வு தேவை.கலைச்சொல் ஆக்கவியலாரின் அடிப்பசைட் செயல்பாடு இலக்கு மொழிக்கு அயல் கருத்துக்களையும் கருத்துருக்களையும் உருப்படுத்தம் செய்ய மிகப் பொருத்தமான அல்லது ஒரளவுக்குப் பொருத்தமான சொற்களைத் தேர்வு செய்வதாகும். இதைச் செய்வதற்குக் கலைச்சொல் ஆக்கவியலார் இலக்கு மொழி பயன்படுத்துபவர்களுக்கிடையில் கலைச்சொற்களின் பொருத்தமான தன்மை, இலக்கணத் தன்மைம, ஏற்றுக்கொள்கை மற்றும் பயன்படு பல்வேறு காரணிகளை மனதில் கொள்ளவேண்டும். முக்கியமான காரணி கலைச்சொற்களின் தரம் ஆகும். இதன்படி மொழியியல் கொள்கைகளின் அடிப்படையில் பொருத்தமான வழிகளைப் பயன்படுத்தி புதிய சொற்கள் உருவாக்கப்டும். மொழிபெயர்ப்பு தரவுத்தொகுதி ஒரு கருத்து, நிகழ்ச்சி, விக்ஷயம் அல்லது கருத்துரு என்பதை உருப்படுத்தம் செய்யப் பலரால் உருவாக்கப்பட்ட பல கலைச்சொற்களின் பெரிய பட்டிலிலிருந்து பொருத்மான ஒன்றைத் தேர்வு செய்யும் முக்கிய பங்களிப்பைச் செய்கின்றது. புதிய சொற்கள் உருவாக்கத்தின் தொடர்ச்சியான முயற்சி தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அமைப்பைத் திட்டமிடுவர்களுக்கு ஒன்றுக்குப் பதிலாக ஒன்றைத் தேர்ந்தெடுப்பதில் சிக்கலைத் தரும். புதியச் சொல்லை உருவாக்குவதா அல்லது இலக்கு மொழியில் பயன்பாட்டால் இயல்பாக்கம் செய்யப்பட்ட மூலமொழிச் சொல்லைப் பயன்படுத்துவதா என்ற விவாதம் எழும். சுப்ல தொழில் நுட்பச் சொற்கள் அவற்றின்

உண்மையான மூலத்தைக் கண்டுபிடிக்க இயலாதவாறு இயல்பாக்கம் செய்யப்பட்டிருக்கும். அன்றாடம் வாழ்க்கையில் அவற்றைப் பயன்படுத்தும் உயர்ந்த நிகழ்வெண் அவற்றைச் சொற்றொகையின் பகுதியாக மாற்றும். எனவே மொழிபெயர் பின் போது அவற்றை இடம் பெயர்க்கத் தேவையில்லை. பொதுவாக இலக்கு மொழியில் பெயர்க்கப்பட்ட தரவுத்தொகுதி மூல மொழியிலிருந்து கடன் வாங்கப்பட்டு புதிய கருத்துக்களையும் கருத்துருக்களையும் வெளிப்படுத்தப் பொருத்தமான சொற்களைத் தேர்ந்தெடுக்க நல்ல மூலவளமாகும்.

5.10.14 சொல் தேர்வு கட்டுப்பாடு

மூல மொழியிலிருந்து தெரிந்தெடுக்கப்பட்ட மிகப் பொருத்தமான நிகரன்காளகக் கருதப்பட வேண்டிய இலக்கு மொழிக்கு சொற்களின் தெரிவு மூலமொழியிலும் பட்டறிவுள்ள மொழியியலாரின் திறமையான மற்றும் கவனமான பொரள்கோளை வேண்டும் மற்றொரு கலவைத்தன்மையான செயல்பாடாகும். சூதன் பொருள் என்னவென்றால் மொழியியலார் இலக்கு மொழியில் இருக்கின்ற ஒரே பொருண்மையத் தருகின்ற வடிவங்களின் பெரிய தொகுப்பிலிருந்து கவனத்தில் கொள்ளவும்.

ஆங்கிலம் : that saint ate food

தமிழ் : அந்த முனிவர் போஜனம் செய்தார்

ஆங்கிலம் : he ate food

தமிழ் : அவன் உணவு சாப்பிட்டான்

மேற்சொன்ன எடுத்துக்காட்டுகளில் நாம் தமிழில் எழுவரின் பண்பின் அடிப்படையில் சரியான இணைச் சொல்லைக் தெரிந்தெடுக்க வேண்டியுள்ளது. மொழியியலாரின் முக்கியமான செயல்பாடு இரு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பில் கருத்தப்படும் வேறுபட்ட சமூக மொழியியல் காரணிகளுக்குப் பொருத்தமான சொல் அலகுகளைக் காண்பதாகும். இம்மாதிரியான எடுத்துக்காட்டுகள் வெற்றிகரமான மற்றும் அறிவு பூர்வமான மொழிபெயர்ப்பில் சொல் தேர்வு கவனமாகக் கையாளப்பட வேண்டும் என்பதைக் காட்டப்படுகினறது. இச்சிக்கல் மனித மொழிபெயர்ப்பில் கவனமாகக் கையாளப்பட்டாலும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பில் புறக்கணிக்கப்படுகினறன. இச்சிக்கலைத் தீர்ப்பதற்கான நல்ல வழி இயந்திரத்தால் படிக்க இயலும் அகராதியின் தனியான இடத்தில் பொருண்மையில் ஒற்றுமையைக் காட்டும் எல்லா

வடிவங்களையும் பட்டியிலிடுவதாகும். இம்மாதிரியான தரவுமையங்களை மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பிரித்தெடுக்க இயலும், பொதுவாக இயந்திரம் படிக்க இயலும் அகராதிகள் பல பாடப்பொருள் களங்களாகப் பிரிக்கப்பட்டிருக்கும். இதனால் ஒரு குறிப்பிட்ட பாடப்பொருளைச் சார்ந்த சொற்களைப் பற்றிய பொருத்தமான தகவலைப் பெற இயலும்.

5.10.15 சொல் மயக்கத்தை நீக்குதல்

இயல்பான சூழலில் மனித கருத்துப் பரிமாற்றம் 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 குறிப்புமரை மயக்கம் (Referential ambiguity) என்றும் தொடரியல் மயக்கம் (sytacitic mbiguity) என்றும் மூன்றாகப் பகுக்கப்படும்.

they went to bank - சொல் மயக்கம்

he love his wife - குறிப்புரை மயக்கம்

he waw a girl in part with a binocular - தொடரியல் மயக்கம்

இயந்திர மொழிபெயர்ப்பு ஒங்கமைப்பு பேசுபவரின் மன உருப்படுத்தத்தின் புலனுணர்வு அடிப்படையில் உருவாக்கப்படுவதால் இது பேசுபவரால் பயன்படுத்தப்படும் சொற்களுக்கும் வாக்கியங்களுக்கும் எல்லைப்படுத்தப் பட்டிருக்கும். இதை நேர் செய்ய மொழிபெயர்ப்பாளர்கள் மூல அகராதியை ஒரு குறிப்பிட்ட சூழலில் பொருத்தமாக் தோன்றுகின்ற நிகரனை இலக்கு அகராதியால் பொருத்துகிறார்கள். சில நேர்வுகளில் இலக்கு மொழியானது மூலமொழியின் சொல்லை உருப்படுத்தம் செய்ய இயலும் நிகரான சொல் அலகைக் கொண்டிருக்காது. இந்நேர்வுகளில் மொழிபெயர்ப்பாளர்கள்ன சொற்றொடரைப் பயன்படுத்தப் முயல்வர் அல்லது நிகரான விளக்கத்தைத் தர முயல்வர்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பை ஆதரிப்பவர்கள் சொல் மயக்கச் சிக்கலை நீக்க இயற்க்கைமொழி உரைகளில் கூடுதலாக நிகழும் மயக்க வடிவுகளில் பெரும் எண்ணிக்கையை ஆய்ந்தும் வெளிப்ப்படையாக உருப்படுத்தம் செய்தும் உயர்ந்த தரமான

மொழிபெயர்ப்பை தரலாம் என்று வாதியிடுகின்றனர். சத்தியம் என்றால் மயக்கமான வடிவுகளும் மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்ட விரிந்த அறிவின் அடிப்படையில் ஆயப்பட்டு இயந்திரத்தால் புரிந்து கொள்ளப்படும் வடிவுகளும் அகராதியில் சேமித்து வைக்கப்பட்ட வேண்டும். ஆனால் பகுத்தறிவாதிகள் இவ்விதம் மொழியியல் தகவல்களைப் பெறும் செயல்பாடு சரியானதும் அல்ல சாத்தியமானதும் அல்ல என்று வாதியிடுகின்றனர். களக்கட்டுப்பாடு மொழிபெயர்ப்பு (domin – specific translation) மொழிபெயர்ப்புக்கு எப்போதும் ஆழமான பொருண்மையியல் ஆய்வு தேவையில்லை என்று வெளிப்படுத்துகின்றன. குறிப்பிட்ட மயக்கம் உள்ள சொல் அலகுகளின் ஆழமான பொருண்மையியல் ஆய்வுக்குப் பதிலாக நேரடியாக மூல மொழியிலிருந்து இலக்கு மொழிக்குப் பதிலீடு செய்யும் அணுகுமுறையான எளிய ஆய்வைத் தெரிந்தெடுப்பது நல்லது. சில சூழல்களில் மயக்கங்கள் மொழி பெயர்ப்பின் போது அப்படியே கொண்டு செல்லப்படும் என்ற எதிர்ப்பு அடிப்படையில் சில சொல் மயக்கங்களைப் புறக்கணித்தல் சாத்தியமானதும் தேவையானதுமாகும். இருப்பினும் சொல் மயக்கங்களின் ஆய்வு, இலக்கு மொழியில் மயக்கமற்ற உருப்படுத்தங்களை உருவாக்கும் என்ற நிலையில் பொதுவான மொழிபெயர்ப்பு நேர்வுகளில் மயக்கத்தைப் புறக்கணிக்கக் கூடாது.

5.10.16 இலக்கணப் பொருத்தம்

மொழியில் மொழிபெயர்ப்பில் இலக்கணப் பொருத்தம் என்பது மூல மொழியில் ஒரு சொல் இலக்கு மொழியில் உள்ள சொல்லுடன் அர்த்தம் உள் மொழிபெயர்ப்பைப் பெற வேண்டி பொருத்தப்படும் மொழிபெயர்ப்பில் பொருத்தத்திற்கு வேறுபட்ட திட்டங்கள் இருக்கின்றன. அவை சொல் நிலைப் பொருத்தம், உருபன் நிலைப் பொருத்தம், இலக்கணப் பொருத்தம், தொடர்நிலைப் பொருத்தம், வாக்கியநிலைப் பொருத்தம் என்பனவாகும். மிகப் பொதுவானது மொழிபெயர்ப்புக்கு எடுத்துக்கொள்ளப்பட்ட இரு மொழிகளுக்கிடையில் வினை பங்களிப்பைப் பொருத்துவதாகும. இலக்கண பொருத்தம் வகைப்பாட்டியியல் அடிப்படையில் வேறுபட்ட இரண்டு மொழிகளுக்கிடையில் செய்யப்படும் இயந்திர மொழி பெயர்ப்பில் முக்கியத்துவம் உள்ளதாகும். தற்போதைய சூழலில் நாம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்ப்பைப் பற்றி பேசும் பொது இது முக்கியத்துவம் வாய்ந்தது: என்னவென்றால் ஆங்கிலத்திற்கு SVO அமைப்பு உண்டு He is food தமிழுக்கு SOV அமைப்பு இருக்கிறது (அவன் சோறு சாப்பிட்டான்). எனவே

இலக்கணப் பொருத்தமும் சொல் வரிசைமுறை மாற்றமும் இலக்கு மொழியில் வெளிப்யீடுகளை வெளிப்டுவதற்கு முக்கியத்துவம் வாய்ந்ததாகும்.

5.11. சுருக்கவுரை

ஒரு தரவுத்தொகுதி ஒற்றை மொழியில் (ஒருமொழி தரவுத்தொகுதி) அல்லது பல மொழிகளில் (பன்மொழி தரவுத்தொகுதி) உரை தரவு இருக்கலாம். பக்கவாட்டு ஒப்பீட்டிற்காக சிறப்பாக வடிவமைக்கப்பட்ட பன்மொழி தரவுத்தொகுதி சீரமைக்கப்பட்ட இணையான தரவுத்தொகுதிகள் என்று அழைக்கப்படுகிறது. இரண்டு மொழிகளில் உரைகளைக் கொண்ட இணையான தரவுத்தொகுதியில் இரண்டு முக்கிய வகைகள் உள்ளன. ஒரு மொழிபெயர்ப்பு தரவுத்தொகுதியில் (translation corpus) ஒரு மொழியில் உள்ள நூல்கள் மற்ற மொழியில் உள்ள உரைகளின் மொழிபெயர்ப்புகளாகும். ஒப்பிடக்கூடிய தரவுத்தொகுதியில், நூல்கள் ஒரே மாதிரியானவை மற்றும் ஒரே உள்ளடக்கத்தை உள்ளடக்கியது, ஆனால் அவை ஒன்றுக்கொன்று மொழிபெயர்ப்புகள் அல்ல. ஒரு இணையான உரையை கையாளுவதற்கு, சமமான உரை பிரிவுகளை (சொற்றொடர்கள் அல்லது வாக்கியங்கள்) அடையாளம் காணும் ஒருவித உரை சீரமைப்பு பகுப்பாய்விற்கு ஒரு முன்தேவை ஆகும் . இரண்டு மொழிகளுக்கு இடையில் மொழிபெயர்ப்பதற்கான இயந்திர மொழிபெயர்ப்பு வழிமுறைகள் பெரும்பாலும் முதல் மொழி கார்பஸையும் இரண்டாவது மொழி கார்பஸையும் உள்ளடக்கிய இணையான துண்டுகளைப் பயன்படுத்தி பயிற்சியளிக்கப்படுகின்றன, இது முதல் மொழி கார்பஸின் உறுப்பு-க்கு-உறுப்பு மொழிபெயர்ப்பாகும்.

மொழியியல் ஆராய்ச்சி செய்வதற்கு தரவுத்தொகுதியை மிகவும் பயனுள்ளதாக மாற்றுவதற்காக, அவை பெரும்பாலும் அடையாளப்படுத்தல் எனப்படும் செயல்முறைக்கு உட்படுத்தப்படுகின்றன. ஒரு தரவுத்தொகுதியைக் குறிப்பதற்கான ஒரு எடுத்துக்காட்டு, பேச்சு-அடையாளப்படுத்தல் அல்லது சொல்வகைப்பாடு அடையாளப்படுத்தல் ஆகும், இதில் ஒவ்வொரு வார்த்தையின் பேச்சின் பகுதியையும் (வினை, பெயர்ச்சொல், பெயரடை, முதலியன) பற்றிய தகவல்கள் தரவுத்தொகுதியில் அடையாளங்களின் வடிவத்தில் சேர்க்கப்படுகின்றன. மற்றொரு எடுத்துக்காட்டு ஒவ்வொரு சொல்லின் சொல்லன் (லெம்மா (lemma) (அடிப்படை (base))

வடிவத்தைக் குறிக்கிறது. தரவுத்தொகுதியின் மொழி அதைப் பயன்படுத்தும் ஆராய்ச்சியாளர்களின் வேலை செய்யும் மொழியாக இல்லாதபோது, சிறுகுறிப்பு இருமொழியாக மாற்றுவதற்கு இன்டர்லீனியர் பளபளப்பு பயன்படுத்தப்படுகிறது.

சில தரவுத்தொகுதிகள் மேலும் கட்டமைக்கப்பட்ட அளவிலான பகுப்பாய்வுகளைப் பயன்படுத்துகின்றன. குறிப்பாக, பல சிறிய தரவுத்தொகுதிகளை முழுமையாக பாகுபடுத்தலாம். இத்தகைய தரவுத்தொகுதிகளை வழக்கமாக கிளையமைப்பு வங்கி (Treebanks) அல்லது பாகுபடுத்தப்பட்ட தரவுத்தொகுதிகள் (parsed corpora) என்று அழைக்கிறார்கள். முழு தரவுத்தொகுதியும் முழுமையாகவும் தொடர்ச்சியாகவும் அடையாளப்படுத்தம் செய்யப்படுவதை உறுதி செய்வதில் உள்ள சிரமம், இந்த தரவுத்தொகுதிகள் பொதுவாக சிறியவை, ஒன்று முதல் மூன்று மில்லியன் சொற்களைக் கொண்டிருக்கும். மொழியியல் கட்டமைக்கப்பட்ட பகுப்பாய்வின் பிற நிலைகள் சாத்தியமாகும், இதில் உருபனியல், பொருண்மையியல் மற்றும் பயன்வழியியல் சிறுகுறிப்புகள் அடங்கும்.

இயல் 6

சொல்வகைப்பாடு அடையாளப்படுத்தல்

6.1 அறிமுகம்

தரவுத்தொகுதிகளையும் உரைகளையும் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்துதல் முக்கியமான செயல்பாடாகும். சொல்வகைப்பாடு அடையாளப்படுத்தப்படாத தரவையும் உரையையும் இயற்கை மொழி ஆய்வுக்குப் பயன்படுத்த இயலாது. குறிப்பாக இயந்திர மொழிப்பெயர்ப்புக்குச் சொல்வகைப்பாடு அடையாளப்படுத்துதல் மிக முக்கியமானதாகும். தரவுத்தொகுதியையும் உரைகளையும் சொல்வகைப்பட்டிறிகு அடையாளப்படுத்தல் பற்றி குறிப்பாக இந்திய மொழியிலிருந்து இந்திய மொழிகளின் மொழிபெயர்ப்பு ஒழுங்குமுறையில் தமிழ்த் தரவுத் தொகுதி எவ்வாறு அடையாளப்படுத்தப்படுகின்றன, எச்சொல்வகை அடையாளங்கள் பயன் படுத்தப்படுகின்றன என்பன குறித்து இங்கு விரிவாக விளக்கப்படும்.

6.2 தரவுத்தொகுதியை அடையாளப்படுத்தல்

தரவுத்தொகுதி பல மொழியியல் தகவல்களுக்காக அடையாளப் படுத்தப்படுகின்றன. சொல்வகைப்பாடு அடையாளப்படுத்தல் (Parts of Speech Annotation), மீக்கூறு அடையாளப்படுத்தல் (Prosodic Annotation), முன்வரு கிளவி அடையாளப்படுத்தல் (Anaphoric Annotation), பொருண்மை அடையாளப்படுத்தல் (Semantic Annotation), கருத்தாடல் அடையாளப்படுத்தல் (Discourse Annotation) என்பன முக்கியமான தரவுத்தொகுதி அடையாளப்படுத்தல்களாகும் (Annotation of corpus).

அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆய்வுக்கு மிகப் பயனுள்ளதாக அமையும். இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி சொற்கள் சொல்வகுப்புகளாக அடையாளப்படுத்தப்பட்டிருக்கும் தரவின் மிகப் பொதுவான வடிவாகும். Brown Corpus, Lorcorpus, BNC என்பவை இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் ஆகும். LLC (Lond Lund Corpus of Spoken English) மீக்கூறுகளுக்காக

அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியாகும். Suffanne corpus தொடரியலுக்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆகும்.

6.2.1 சொல்வகைப்பாடு அடையாளப்படுத்தல்

சொல்வகைப்பாடு அடையாளப்படுத்தலின் (Parts of Speech Annotation) நோக்கம் ஒரு பனுவலில் உள்ள ஒவ்வொரு சொல் அலகுக்கும் அதன் வகைப்பாட்டைக் காட்டும் ஒரு கூறிலே தருவதாகும். இது விரிதரவிலிருந்து தரவை மீட்டெடுப்பதில் தனித் தன்மையைக் கூட்டுகிறது. இது தொடரியல் பகுத்து அடையாளப்படுத்தலுக்கும், பொருண்மைக்கள அடையாளப்படுத்தலுக்கும் உதவுகிறது. இது ஒற்றுமையான எழுத்துக்களை வேறுபடுத்துவதற்கு நம்மை அனுமதிக்கின்றது. சொல்வகை அடையாளப்படுத்தல் பின்வரும் தன்மைகள் கொண்டதாகும்:

- ஒவ்வொரு சொல்லுக்கும் ஒரு சொல்வகைப்பாட்டை அல்லது பிற சொல்வகை அடையாளத்தைத் தருவதாகும்.
- அடையாளப் படுத்துகைப் பொதுவாக நிறுத்தற்குறிகளுக்கும் பயன்படுத்தப்படும்.
- கணினிமொழியில் அடையாளப் படுத்துதல் [tokenization] போல் தான் இயற்கை மொழியிலும் சொல்லை அடையாளப்படுத்தும் செயல்பாடு அமையும்.
- சொல்வகை அடையாளப்படுத்தல் பேச்சுப் புரிந்து கொள்ளுதல் [Speech Recongition] இயற்கை மொழி பகுத்துக் குறித்தல் [Natural Language Parsing]மற்றும் தகவல் மீட்டுப் பெறல் [Information retrival] என்பனவற்றில் முக்கியமான பங்கு வகிக்கின்றது.
- சொல்வகை அடையாளப்படுத்தலின் உள்ளீடு [Input] வழிமுறை வரைவு [algorithm] சொற்களின் கோர்வைகளும் குறிப்பிட்ட சொல்வகைப்பாட்டு அடையாளக் குழுமமும் [Part of Speech Tagsets] ஆகும்.
- விடுவரல் [Output] ஒவ்வொரு சொல்லுக்கும் ஒவ்வொரு நல்ல அடையாளத்தைத் தருவதுதான்.

எ.கா:

அவன் <மா.பெ> என் <வ.மா.பெ> சொல் <வி.பெ> படி <வி.பெ.பி.உ> கேட்டான்.

POS அடையாளப்படுத்தியின் முக்கியமான வேலை சூழலுக்குத் தக்கவாறு சரியான அடையாளத்தைத் தேர்ந்தெடுத்து மயக்கத்தைத் தீர்ப்பதுதான். பெரும்பாலான தமிழ்ச் சொற்கள் எல்லாம் மயக்கம் இல்லாதவை. இருப்பினும் பொதுவாகப் பயன்படுத்தப்படும் சொற்கள் மயக்கம் உள்ளவை.

6.2.2 முன்வருகிளவி அடையாளப்படுத்தல்

முன்வருகிளவி அடையாளப்படுத்தலில் (Annaphic Annotation) எல்லா மாற்றுப் பெயர்களும் பெயர்த்தொடர்களும் அகப்பற்றாக (Cohesion) என்பதன் விரிவான சட்டகத்திற்குள் ஒரே குறிபடு பொருளைக் குறிப்பதற்காக அடையாளப்படுத்தப்படுகிறது. இங்கு முன்வருகிளவியும் வேறுபட்ட வகைகள் வகைப்படுத்தப்பட்டு வரிசைப்படுத்தப்படுகின்றன. இந்த முன்வருகிளவி அடையாளப்படுத்தப்படுத்தல் திட்டம் மாற்றுப்பெயர் தீர்வு (Anaphora resolution) போன்ற ஆய்வு மற்றும் பரிசோதனைச் செய்யு[ம் இயக்க நெறிமுறைகளுக்குப் பயன்படுத்தப்படுகின்றது.

எ.கா.

ரவி தன் தங்கையைப் பார்க்கப்போகும் ராமுவைச் சந்தித்தான்

இது பனுவல் புரிந்து கொள்கைக்கும், இயந்திர மொழிபெயர்ப்புக்கும் முக்கியமானதாகும்.

6.2.3 மீக்கூறு அடையாளப்படுத்தல்

மீக்கூறு அடையாளப்படுத்தலின் (Prosodic Annotation) இசையோட்ட ஒழுங்மைப்புகள், அழுத்தம் மற்றும் விட்டிசை என்பன பேச்சில் அடையாளப் படுத்தப்படும் இது மிகச் சிரமமான அடையாளப்படுத்தலே ஆகும். ஏனெனில் மீக்கூறு பிறமொழி நிலைகளை விட இயல்பில் தன் உணர்ச்சி அடிப்படையில் அமையும். இது பயிற்சி பெற்ற காதால் கவனமாகக் கேட்கப்படுத்தலைக் கேட்கும் Lancaster/IBM Spoken English Corpus மீக்கூறுக்காக அடையாளப்படுத்தப்பட்ட விரிதரவாகும். இதில் அழுத்தப்பட்ட அசைகள் சுதந்திரமான இசைமை இயக்ககத்துடன் மற்றும் இயக்கம் இல்லாமல் வேறுபட்ட குறியீடுகளால் அடையாளப்படுத்தப் பட்டுள்ளன. இசைமையைச் சூழலிலிருந்து ஊகிக்க இயலும். அழுத்தம் இல்லாத அசைகள் அடையாளப்படுத்தப்படாமல் விடப்பட்டுள்ளன.

6.2.4 பொருண்மை அடையாளப்படுத்தல்

பொருண்மை அடையாளப்படுத்தலில் (Semantic Annotation) பனுவலில் உள்ள சொற்களின் பொருண்மைப் பண்புக் கூறுகளோ (குறிப்பாக சொல் அர்த்தத்தின் அடையாளப்படுத்துகை) அல்லது சொற்களுக்கு இடையேயுள்ள பொருண்மை உறவுகளோ (குறிப்பிட்ட செயலின் செயல்கள், பாதிப்பவர்கள் போன்றவை) அடையாளப்படுத்தப்படுகின்றது. எ.கா.

அவள் பழம் சாப்பிட்டாள்

அவள் - (படர்க்கை பெண்பால், ஒருமை, மாற்றுப்பெயர்,பதிலீடுபெயர்

பழம் - (தாவரத்தின் பூவிலிருந்து விளையும் விதையுள்ள இனிப்பான பொருள்

சாப்பிட்டாள் - (பழத்தை உட்கொள்ளுதல்)

எந்தப் பொருண்மைப் பண்புக் கூறுகள் அடையாளப்படுத்தப்பட வேண்டும் என்பதில் ஒருங்கிணைந்த கருத்து இல்லை. சிலர் Roget's Thesaurus –இல் உள்ள பொதுமையான பொருண்மை வகைப்பாடுகளை முன் வைத்தனர். இம்மாதிரியான அடையாளப்படுத்தும் திட்டம் சொற்களின் மூடப்பட்ட வகுப்புகளையும் திறந்த வகுப்புகளையும் நடைமுறைப்படுத்த அடையாளப்படுத்தப் படுகின்றது.

6.2.5 கருத்தாடல் அடையாளப்படுத்தல்

கருத்தாடல் அடையாளப்படுத்தலில் (Discourse Annotation) ஒரு தரவுத்தொகுதி பனுவல் மற்றும் கருத்தாடல் நிலைகளில் அடையாளப்படுத்தப்படுகின்றது. அடையாளப்படுத்துகின்ற வகை கருத்தாடல் ஆய்வில் முக்கியப் பங்களிப்பைச் செய்தாலும் பரவலாகப் பயன்படுத்தப்படுவதில்லை. ஏனெனில் மொழி வகைப்பாடுகள் சூழல் சார்ந்தவை மற்றும் அவற்றைப் பனுவல்கிளல் அடையாளம் காணுதல் பிற மொழியியல் நடப்புகளைக் காட்டிலும் விவாதத்திற்கு கூடுதல் மூல காரணமாக அமையும். Londond – Lund – Spoken Corpus–இல் வயதுக்கு வந்தவர்களின் பேச்சு போக்கைக் கவனிக்க 16 கருத்தாடல்கள் பயன்படுகின்றது.

6.2.6 பகுப்பாய்வு

பகுப்பாய்வு என்பது ஒரு இலக்கண அடிப்படையில் பனுவல்களின் தானியக்க ஆய்வுகளுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு பனுவலுக்கு தொடரியல் அமைப்பைத் தரும் செற்பாங்கைக் குறிப்பிடுகின்றது. இது பெரும்பாலும் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுப்பிடிக்கப்பட்ட பின் நடைமுறைப்படுத்தப்படும்.

வேறுபட்ட இலக்கணங்கள் (சார்பு இலக்கணம், சூழல் இலக்கணம் போன்றவை) அடிப்படையில் பகுத்துக் குறித்தல் ஒன்றுடன் ஒன்று உயர்ந்த நிலைத் தொடரியல் உறவுகளில் கொண்டு வருகிறது. வாக்கிய நிலைப் பகுத்துக் குறித்தல் சொல்நிலைப் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்த தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உள்ளடக்குகின்றது. ஒரு பகுத்து அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பகுத்துக் குறித்தலில் கிளைப்படங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree bank) என்றழைக்கப்படுகின்றது. கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் அடையாளப்படுத்தப்பட்ட அடைப்புக்குறிகளின் குழுமத்தைப் பயன்படுத்தி உருப்படுத்தம் செய்யப்படும்.

எ.கா:

'Pearl sat on a chair' என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்.

[S [NP Pearl – NP/NP] [VP – sat VVI] [PP on II]

[NP a – ATI] [chair – NNI]

[NP] [PP] [VP] S]

இதில் உருபனியல், தொடரியல் தகவல் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்பட்டிருக்கும் உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் திறக்கும் மற்றும் அடைக்கும் செவ்வக அடைப்புக் குறிகளால் அடையாளப்படுத்தப்படும்.

எல்லாப் பகுத்துக் குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருமாறு:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை.

2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வகைகள்

இந்த வேறுபாடுகள் இருப்பினும் பெரும்பாலான பதுத்துக் குறித்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையால் அமைந்தது.

இந்த அமைப்பொழுங்குக்குள் முழு பகுத்துக் குறித்தல் [full parsing] திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாகக் கொண்டது. குறுகிய பகுத்துக் குறித்தல் திட்டம் (Skeleton parsing scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தலைப் பயன்படுத்துகின்றது. மேலும் சில உறுப்பு வகைகளின் அக அமைப்பை

விட்டுவிடுகின்றது. பகுத்துக் குறித்தல் பெரும்பாலும் மனித ஆய்வில் பின் திருத்தம் செய்யப்படுகின்றது. ஏனென்றால் தானியங்கு பகுத்துக் குறித்தல் சொல்வகை அடையாளப்படுத்தலை விட வெற்றி குறைந்தது. முழு மனித இயக்கப் பகுத்துக் குறித்தலின் குறைப்பாடு பகுத்துக் குறித்தல், தரவுத்தொகுதி திரித்தல் இவற்றில் ஈடுபடும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுக்கட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள் கோள் சாத்தியமாகும்போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் வேறுபட்ட நிலைகளில் (சொல்நிலை, வாக்கியநிலை, தொடர்நிலை, செயல்பாடு–பங்கெடுப்பாளர் நிலை) அடையாளப்படுத்தும் நிலை தரும் மொழி மூலவளமாக அமைகின்றது. கிளை வங்கிகள் இயற்கைமொழி ஆய்வு தரவு உந்தல் அடிப்படைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுத்தல், பொதுவாக மொழியியல் ஆய்வுகள் இவற்றின் உருவாக்கத்திற்கு முக்கியமானதாகும்.

6.2.7 தலைச்சொல்லாக்கம்

தலைச்சொல்லாக்கம் (Lemmatization) செயல்பாடு விரிதரவில் பயன்படுத்தப்பட்டுள்ள சொற்களின் மூல வகைப்பாட்டைக் கண்டுபிடிப்பதும் சொற்களை அவற்றுடன் தொடர்புடைய சொல்லன்களாகக் (Lexeme) குறைப்பதும் ஆகும். எடுத்துக்காட்டாக *படித்தான்*, *படிக்கிறான்*, *படிப்பான்*, *படிப்பாள்* என்ற சொல் வடிவுகளிலிருந்து *படி* எனக் குறைத்தல். இது சொல்லின் எல்லா மாற்று வடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச் சொல்லின் மாற்றுவடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச்சொல்லின் மாற்று வடிவங்களைப் பிரித்தெடுக்கவும், பரிசோதிக்கவும் ஆய்வாளர்களை அனுமதிக்கின்றது. இது மொழி கற்றலுக்குப் பயன் உள்ளதாக அமைகின்றது. இங்கு கற்பவர்கள் ஒரு தலைப்புச் சொல்லின் மொத்த சாத்தியமான எண்ணிக்கையில் பயிற்சி பெறுகின்றார்கள். இது எந்தச் சொற்கள் திரிபுறுகின்றன என்பதை அறிவதற்குப் பயன் உள்ளதாக இருக்கின்றது. எடுத்துக்காட்டாக Brown Corpus–இன் ஒரு பகுதி சொல் மற்றும் இலக்கணத் தகவலுடன் கூடிய சொற்களின் தலைப்பாக்கம் செய்யப்பட்ட வடிவுகளைக் கொண்டிருக்கின்றது.

6.3 சொல்வகை அடையாளப்படுத்தலின் முக்கியத்துவம்

மொழி ஆய்வில் சொல் வகைப் பாட்டின் முக்கியத்துவம் என்னவென்றால் இது சொல்லைப் பற்றியும் அவற்றை அடுத்து வரும் சொற்கள் பற்றியும் செய்திகளை ஏராளமாகத் தருகின்றது. இவை முக்கிய வகைப்பாடுகளுக்கு உண்மையாகும். வினை-பெயர் மட்டுமின்றி பல நுட்பமான வேறுபாடுகளுக்கும் உண்மையாகும். எடுத்துக்காட்டாக இந்த அடையாளக் குழுமங்கள் உடமை மற்றும் பெயர்களுக்கும் (*என்னுடைய, உன்னுடைய, அவனுடைய, அவளுடைய, அவருடைய, அதனுடைய*) மூவிடப் பெயர்களுக்கும் (*நான், நீ, அவன், அவள், அது*) வேறுபாட்டைக் காட்டுகின்றன.

ஒரு சொல் உடமை மாற்றுப்பெயரா? அல்லது மூவிடப்பெயரா? என்பதை அறிந்துகொண்டு எந்தச் சொற்கள் அதன் பக்கத்தில் வர இயலும் என்பதைக் கூற இயலும். இதைப் பேச்சுத் தெரிந்துகொள்ளுதல் (Speech Recognition) என்பதற்கான மொழி மாதிரியில் (Language Model) பயன்படுத்தலாம். சொல்வகைப்பாடு தகவல் மீட்டலுக்கான (Information retrieval) வேர்ச்சொல்லைக் கண்டுபிடித்தலுக்குப் பயன்படுகின்றது.

ஏனென்றால் சொல்வகைப்பாட்டை (Parts of Speech) அறிவது அவை எந்த உருபனியல் ஒட்டுகளையும் ஏற்கும் என்பதை அறியலாம். அவை பனுவலிலிருந்து பெயர்களையோ முக்கிய சொற்களையோ அறிய உதவுவதால் தகவல் மீட்டல் பயன்பாட்டிலும் பயன்படுத்தலாம். தானியங்கு சொல்வகைப்பாட்டு அடையாளப் படுத்திகள் சொல்பொருள் மயக்கம் நீக்கி வழியமைப்பு முறை (word sense disambiguation algorithms) சொல் வகைப்பாட்டு அடையாளப்படுத்திகள் (Part-of-speech-taggers=pos taggers), வகுப்பு அடையாளப்படுத்திகள் (class based N-grams) போன்ற முன்னேறிய ASR மொழி மாதிரிகளில் பயன்படுகின்றது.

ஆங்கிலத்தில் புகழ்பெற்ற POS tagger Brown Corpus-இல் பயன்படுத்தப்பட்ட 87 அடையாளக் குழுமங்களிலிருந்து (tagsets) உருவாக்கப்பட்டவை. மூன்று மிகப் பொதுவாகப் பயன்படுத்தப் படுகின்றன.

1. 45--Tag Penn Tree bank tagset.

2. The medium Size 61 tag 65 tagsets used by Lancaster VCREL, Projects CLAWS (Constituent Liklihood Audomatic Word-Tagging System) Tagger to tag British National Corpus (BNC)

3. The Longer 146-tag C7 tagset

Ex.

Penn Tree bank Version of Brown Corpus

ASCII Flat file

Tag represented after each word following a slash

The/DT grand /JJ/ Jury/NN commented/VBD on/IN a/DT number/NN of/IN others/JJ Topics/NNS%

DT – Deter miners

JJ – Adjective

NN – Noun,sing,or mass

NBD – Verb, Past tense

IN – Preposition

NNS – Noun plural

6.4 சொல் வகைப்பாட்டை அடையாளப்படுத்தலின் வகைகள்

மூல சொல்வகை அடையாளப்படுத்துவதின் வழிமுறை வரைவு (Algorithm) இரு வகுப்புகளில் அடங்கும். மூன்று அடையாளப்படுத்திகள் (taggers) பயன்பாட்டில் உள்ளன. அவையாவன:

1. விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Rules based taggers)
2. புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Stochatic Taggers)
3. மாற்றம் அடிப்படையிலான அடையாளப்படுத்திகள் (Transformation based taggers)

6.4.1 விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்

விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் மயக்கம் தருகின்ற சொல் பெயரா?, வினையா? என்று குறிப்பிடுகையால் எழுதிய பெரிய தரவு மையங்களை [data base] உள்ளடக்கியது. ஆரம்பக் கட்டத்தில் தானாகவே சொல்வகைப்பாட்டை அடையாளப்படுத்துகின்றது. அவ்வாறு அடையாளப்படுத்த பயன்படுத்தப்பட்ட வழிமுறை வரைவுகள் (Algorithm) இரண்டு நிலை அமைப்பால் ஆனதாகும்.

1. அகராதி ஒவ்வொரு சொல்லுக்கும் சாத்தியமான சொல்வகைப்பாட்டைத் தரும்.

2. இரண்டாம் நிலை கையால் எழுதப்பட்ட மயக்கம் தரும் சொல்வகைப்பாட்டைத் தருமாறு பயன்படுத்தும்.

எடுத்துக்காட்டாக ENG Two Level Tagger இரண்டாம் நிலை அமைப்பால் ஆனதாகும். ENG TWO Level அகராதி இருநிலை உருபனியல் [Two level Morphology] அடிப்படையில் அமைந்தது.

6.4.2 புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Stochostic Part of Speech Tagging)

சாத்தியங்கள் அடையாளப்படுத்துவதில் பயன்படுத்துவது மிகப் பழமையானது. பல சாத்திய அடிப்படையிலான அடையாளப்படுத்திகள் வழக்கில் உள்ளன. எ.கா. Hidden Markove Model or HMM Tagger. இந்தப் புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தியில் பயன்படுத்தப்படும் உள்ளுணர்வு 'ஒரு சொல்லின் கூடுதலான சாத்தியமான அடையாளத்தை எடு' [Pick up the most likely tag for the Word] என்பதாகும். தரப்பட்ட வாக்கியத்திற்கு அல்லது சொல் கோர்வைக்கு HMM அடையாளப்படுத்தி பின்வரும் வாய்ப்பாட்டைப் (formula) பெரிதுபடுத்தும் அடையாளக் கோர்வையைத் தேர்ந்தெடுக்கின்றது.

P(wird/Tag) *P(Tag/Previous n tag)

HMM ஒரு சொல்லைக் காட்டிலும் ஒரு முழு வாக்கியத்திற்கு அடையாளக் கோர்வையைத் தருகின்றது. இவ்வகையான Bigram – HMM Tagger Wi என்ற சொல்லுக்கு ti என்ற அடையாளத்தைத் தேர்ந்தெடுக்கும். அது முந்தைய அடையாளம் ti–ஐ மற்றும் தற்போதைய சொல் wi என்று தரப்படுகையில் கூடுதலான சாத்தியமான அடையாளத்தைத் தேர்ந்தெடுக்கும்.

$$ti = \arg\max_{j} P(tj/ti\text{-}l, wi)$$

ஒரு அடையாளப் படுத்தலுக்கு அடிப்படையான HMM சமன்பாடு

$$ti \; \text{argmax}_{j} \; (tj, Pi\text{-}l) \; P(witj)$$

எ.கா:

(1) Secretariat,NNP is /VB2 expected/VBN to/To race VB Tomorrow/NN (2) people/NNs continue NBP to/To inquire/VB The DT Reason/NN for/in The/CT race/NN For/in outer/JJ Space/NN.

முதல் எடுத்துக்காட்டில் race ஒரு வினை (VB) இரண்டாவது எடுத்துக்காட்டில் race ஒரு பெயர் (NN) Bigram HMM அடையாளப்படுத்தி அடையாளச் சிக்கலை அண்மையிலுள்ள சொற்களையும், அடையாளங்களையும் பார்த்துத் தீர்மானிக்கலாம் என்று ஊகம் செய்கின்றது.

To/To race/?????

Tge/DT race/????

எவ்வாறு race என்ற நமது எடுத்துக்காட்டு பயன்படுகின்றது என்று பார்ப்போம். நாம் NN -க்கும் VB -க்கும் இடையே ஒன்றைத் தேர்ந்தெடுக்க வேண்டும். நாம் எதற்குக் கூடுதல் சாத்தியம் இருக்கின்றதோ அதைத் தேர்ந்தெடுக்க வேண்டும்.

P(VB/To) P(race/VB)

P(NN/TO) P(race/NN)

பின்வருவது சாத்தியங்களாகும் (probabilities)

P(NN/TO) = :021

P(VB/To) = .34

6.4.3 மாற்றம் அடிப்படையிலான அடையாளப்படுத்திகள்

மாற்றம் அடிப்படையிலான அடையாளப்படுத்துதல் (Transformation Based Tagging) Bill Tagging என்று அழைக்கப் படுகின்றது. இது பொறி கற்றலுக்கான மாற்றம் அடிப்படையிலான கற்றல் (Transformation based Learning – TBL) அணுகுமுறையாகும். இது விதி அடிப்படையிலிருந்தும் புள்ளியியல் அடிப்படையிலிருந்தும் செய்திகளை எடுக்கின்றது. விதி அடிப்படையிலான அடையாளப்படுத்திகளைப் போல் TBL உம். எந்தச் சொற்களுக்கு எந்த அடையாளத்தைத் தரவேண்டும் என்று குறிப்பிடுகின்றது. ஆனால் அடிப்படையிலான அடையாளப் படுத்திகளைப் போல் TBL ஒரு மொழி கற்றல் நுணுக்கமாகும் (Machine Learning Technique). இதில் தரவிலிருந்து விதிகள் தானாகவே ஊகிக்கப் படுகின்றது. சில HMM- Taggers -ஐ போல் TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் நுணுக்கமாகும் , (Supervised Learning Technique). இது ஏற்கெனவே அடையாளப் படுத்தப்பட்ட பயிற்சி பெற்ற பெருந்தரவை [ore-

Tagged Training Corpus] எதிர்ப்பார்கின்றது,ஊகிக்கின்றது. TBL -உக்கு ஒரு குழுமம் அடையாளப்படுத்தும் விதிகள் இருக்கின்றது. முதலில் பெருந்தரவு பெரும்பாலும் பயன்படுத்தப்படும் விரிந்த விதிகளை வைத்து அடையாளப்படுத்தப்படும். பின்னர் சில முன்னதான [original] அடையாளங்களை மாற்றும். மேலும் திட்டவட்டமான விதிகள் தேர்ந்தெடுக்கப்படும். பின்னர் குறைவான எண்ணிக்கையிலான அடையாளங்களை மாற்றும். மேலும் திட்டவட்டமான விதிகள் பயன்படுத்தப்படும். எவ்வாறு TBL விதிகள் பயன்படுத்தப்படுகின்றது (How TBL Rules-are applied) என்பது கீழே விளக்கப்பட்டுள்ளது.

Brill's Tagger – ரை கொண்டு பயன்படுத்தப்படும் விதிகள் நடைமுறையைப் பயன்படுத்துவதற்கு முன் அடையாளப்படுத்தி ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும் இந்த கூடுதல் அடையாளங்களை நாம் பெருந்தரவில் பெறலாம்.

Brown Corpus –இல் 'race' என்பதற்குக் கூடுதல் உகந்தது பெயர் என்பதாகும்.

P(NN/race)=.98

P(VB/race)=.02

நாம் முன்னர் பார்த்த race என்பதன் இரு எடுத்துக்காட்டுகளும் NN என்று அடையாளப்படுத்தப்பட்டிருக்கும். முதல் எடுத்துக்காட்டில் NN என்பது தவறான அடையாளம்.

is/VB2 expected/VBN to/To race/NN tomorrow

இரண்டாவது எடுத்துக்காட்டில் race என்பது சரியாக NN என்று அடையாளப்படுத்தப்பட்டிருக்கும்.

The/DT race/NN for/IN outer/JJ space/NN

கூடுதல் உகந்த அடையாளத்தைத் தேர்ந்தெடுத்தப்பின் சேடைடள கூயபபநச அதன் மாற்ற விதிகளை நடைமுறைப் படுத்தும். இது நடக்கையில் race என்பதன் தவறான அடையாளப்படுத்தலில் நடைமுறைப்படுத்தப்படும் விதியைக் கற்றிருக்கும். முன்வரும் tag to என்றால் NN என்பதை VB என்றும் மாற்றுவதற்து முன் to/To என்று வந்தால் இந்த விதி race/NN என்பதை race/VB என்று மாற்றும்.

Expected/VBN to/To race/NN

Expected/VBN to/To race/VB

எவ்வாறு TBL விதிகள் கற்கப்படுகின்றன என்பது பின்வருமாறு விளக்கப்படும். Brill's TBL வழிமுறை வரைவுக்கு [algorithm] மூன்று முக்கிய நிலைகள் இருக்கின்றன. இது முதலில் ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும். பின்னர் இது ஒவ்வொரு சாத்தியமான மாற்றத்தையும் (Possible Transformation) பரிசோதிக்கும். இறுதியில் இருந்த விதிப்படி தரவை மீள் அடையாளப்படுத்தல் செய்யும். முன்னால் வரும் அடையாளப்படுத்தலை விட குறைந்த மேம்பாடுடைய அடையாளப்படுத்தலான நிறுத்தல் அளவீடு வருவது வரை இந்த மூன்று நிலைகளும் திரும்பித்திரும்பி சொல்லின் சரியான அடையாளம் தெரியும் என்பதை வேண்டுகிறது. அதாவது TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் வழிமுறை வரைவு (Supervised learning algorithm) ஆகும்.

TBL செயல்பாட்டின் வெளிப்யீடு முறைப்படுத்தப்பட்ட மாற்றங்களின் பட்டியலாகும். இவை புதிய பெருந்தரவுக்குப் பயன்படுத்தத்தக்க,நடைமுறை படுத்தத்தக்க அடையாளப்படுத்தல் செயல் முறையை (tagging Procedure) உருவாக்கும். கொள்கை அடிப்படையில் சாத்தியமான மாற்றங்களின் குழுமம் எல்லையற்றதாகும். ஆனால் மாற்றங்கள் ஒவ்வொன்றும் வழிமுறை வரைவைக் கடக்குமாறு மிகப் பொருத்தமான ஒன்றைத் தேர்ந்தெடுக்க TBL ஒவ்வொரு மாற்றத்தையும் எடுத்துக்கொள்ள வேண்டும். இவ்வாறு மாற்றங்களின் குழுமத்தை எல்லைப்படுத்த வழிமுறை வரைவுக்கு ஒரு வழி தேவை. இது சிறிய templates -களின் குழுமத்தை (உருமான மாற்றங்களை) உருவாக்கிச் செய்யப்படுகின்றது. பின்வருவன Brill's Templetes -களின் குழுமத்தின் பட்டியலாகும். முன்னர் வருகிற (தொடர்நது வருகிற) சொல் Z-ஆக அடையாளப்படுத்தப்படும். இரண்டு சொற்களுக்கு முன்னால் (பின்னர்) வருகிற சொல் Z -ஆக அடையாளப்படுத்தப்படும்.

முன்னர் வருகிற இரு சொற்களில் ஒன்று Z -ஆக அடையாளப் படுத்தப்படும். முன்னர் வருகிற மூன்று சொற்களில் ஒன்று Z -ஆக அடையாளப் படுத்தப்படும். முன்னர் வருகிற சொல் Z -ஆகவும் இரு சொற்களுக்கு முன்னால் (பின்னால்) வரும் சொல் W -ஆகவும் அடையாளப்படுத்தப்படும்.

6.5 சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம்

6.5.1 சொல்வகை அடையாளக் குழுமங்கள் பற்றிய முந்தைய முயற்சிகள்

6.5.1.1 தொடக்ககாலச் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம் [1980-க்கு முன்]

1960-களிலும் 1970-களிலும் ஆங்கில மொழிக்கு வேண்டி சொல்வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கும் செயல்பாடு ஐக்கிய அமெரிக்காவில் தொடங்கப்பட்டு விட்டது. ஆரம்பக் காலத்தின் முக்கியமான சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் Klein – Simmons(1963) மற்றும் Gree – Fubin (1971) என்பவர்களால் உருவாக்கப்பட்டது. மற்றொரு முயற்சி பென்சிலியா பல்கலைக்கழகத்தில் மேற்கொள்ளப்பட்டது.

6.5.1.2 ஆங்கில சொல் வகைப்பாட்டுக் குழுமங்கள் (1980-க்குப் பிறகு)

1980-1990-களில் லங்காஸ்டர் பல்கலைக்கழகத்தில் ஆங்கிலத்திற்கு வேண்டி ஒரு தொடர்ச்சியான சொல்வகைப்பாட்டு குழுமங்கள் CLAWS என்ற சொல்வகைப்பாட்டை அடையாளப்படுத்தும் மென்பொருளை பயன்படுத்தி உருவாக்கப்பட்டது. CLAWS என்ற சொல்வகைப்பாட்டு அடையாளக் குழுமம் LOB தரவுத்தொகுதிக்குப் பயன்படுத்தப்பட்டது. பின்னர் CLAWS -ன் C5 மற்றும் C7 சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் BNC விரிதரவை அடையாளப் படுத்தப் பயன்படுத்தப்பட்டது. இதைத் தொடர்ந்து பல ஆங்கில சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் பயன்படுத்தப்பட்டன. அவையாவன:

1.The TOSCA Scheme

2.The ICE Tagset

3.The PENN Tagset

4.The LUND Tagset

5.The Eng CG Tagset

6.5.2 அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு

ஆங்கிலம் தவிர பிற மொழிகளுக்குச் சொல் வகைப்பாடு அடையாளம் செய்யும் போது சொல் வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்தும் தேவை வெளிப்பட்டது. சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் மிக சமீபயத்திய தரமான ஒன்று EAGLES— வழிகாட்டல்களாகும் (The ENGLES-Guidelines) [Leeach and Wilson] இதனுடைய நோக்கம்

பலவிதமான திட்டங்களுக்கும் பல மொழிகளுக்கும் சொல் வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்துவதாகும்.

EAGLES வழிகாட்டல்கள் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் உள்ளடக்கும் அல்லது உள்ளடக்க வேண்டிய ஒரு குழும பண்புக்கூறுகளைத் தருகிறது. இதனுடன் இந்த எல்லாப் பண்புக்கூறுகளையும் குறியாக்கம் செய்து இடைநிலைப்பட்ட குறியாக்கத்தின் தேவை என்னவென்றால் இரண்டு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கு இடையில் மொழிப்பபெயர்ப்பை அனுமதிப்பதாகும். இதனால் அவற்றிற்கிடையில் பொருத்தத்தை உறுதி செய்வதாகும். உருப-தொடரியல் அடையாளப்படுத்தலுக்கு EAGLES வழிகாட்டல்களில் மூன்று காட்டல்களில் அமைப்பாக்கம் செய்யப்பட்டுள்ளது.

1. எவை எவற்றைக் கட்டாயமாகக் கருதப்படுகிறது.

2. எவை பரிந்துரைக்கப்படுகின்றது.

3. எவை உருப-தொடரியல் அடையாளப் படுத்தலுக்கு சின்னம் - மதிப்பிணைகளாக வரையறை விளக்கம் செய்யப்படுகிறது. எடுத்துக்காட்டாக 'பால்' என்ற சின்னத்திற்கு EAGLES -இன் பரிந்துரைகளில் ஆண்,பெண் அல்லது பால் மதிப்பிணைகள் படிநிலையாக அமைக்கப்படவோ அமைக்கப்படாமலோ இருக்கலாம்.

EAGLES வழிகாட்டல்களால் எந்தச் சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கும் கட்டாயமாக பதிமூன்று பரிந்துரைக்கப்படுகின்றது.

1. Noun (பெயர்)
2. Verb (வினை)
3. Adjective (பெயரடை)
4. Pronoun (பதிலீடு பெயர்)
5. Article (சுட்டிடைச்சொல்)
6. Adposition (உருபு)
7. Adverb (வினையடை)
8. Conjunction (இணைப்பான்கள்)
9. Numeral (எண்:
10. Interjection (வியப்பிடைச்சொல்)

11. Unassigned/unique (இணையற்ற,தனிப்பட்ட)

12. Residual (மீதியான)

13. Punctutation (நிறுத்தற்குறிகள்)

பரிந்துரைக்கப்பட்ட மற்றும் இருப்புச் சின்னங்கள் இந்த முக்கியமான சொல் வகைப்பாடுகளால் ஒழுங்குப்படுத்தப் பட்டுள்ளது. சொல் வகைகளை கடந்து பொருத்தமுறத் தேவையில்லை.

EAGLES வழிகாட்டல்கள் சொல் வகைப்பாட்டு அடையாளக் குடும்பத்தைத் திட்டமிடுபவரின் சுதந்திரத்தைக் கட்டுப்படுத்தாமல் அவர் குறிப்பிட்ட எண்ணும் எல்லா அடிப்படையான செய்திகளையும் உள்ளடக்கி நிகழ இயலும் சட்டகத்தைத் தருகின்றது. முன்னர் விளக்கப்பட்ட பல அடையாளக் குழுமங்கள் EAGLES வழிகாட்டல்களில் எளிதில் பொருந்தாத பண்புக் கூறுகளை உள்ளடக்கலாம். இருப்பினும் இவைகள் மற்றொரு அடையாளக் குழுமத்திற்கு மொழிப்பெயர்க்க இயலாத குழுமமும் அவற்றை உள்ளடக்கவில்லை. எனவே அவற்றை EAGLES இடைப்பட்ட அடையாளக் குழுமத்தில் அடக்குவது தேவையில்லாததாகும். இந்தச் சட்டகத்தின் மதிப்பு இது வேறுபட்ட மொழிகளின் மொழியியல் வளங்களின் திண்மையையும் திரும்ப பயன்படுத்தலையும் ஊக்குவிக்கின்றது. மற்றும் சுழலான திரும்பக் கண்டுப்பிடித்தலை ஊக்கமிழக்கச் செய்கிறது.

EAGLES வழிகாட்டலின் முக்கியமான குறை இது. உலக மொழிகளின் மிகக் குறுகிய பகுதியைதான் அடக்குகிறது. ஐரோப்பிய ஒருங்கிணைப்பின் திட்டமாக இது ஆங்கிலம், டச்சு, ஜெர்மனி, டேனிசுஷ், பிரெஞ்சு, ஸ்பேனிசுஷ், போர்த்துகீசு, கிரீக், இட்டாலின் என்ற வகைப்பாட்டியல் அடிப்படையில் ஒற்றுமையுள்ள ஒன்பது மொழிகளைத் தான் உள்ளடக்குகிறது. எனவே இவை அல்லாத பிறமொழிகளுக்கு நுஆழுடுநுளு வழிகாட்டல்களின் பயன்பாடு பொருத்தமாகவோ, பொருத்தமற்றதாகவோ இருக்கலாம்.

6.5.3 EAGLES வழிகாட்டல்களின் அடிப்படையில் உருவாக்கப்பட்ட சில சமீபக் காலத்திய சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

EAGLES வழிகாட்டலின் வெளிப்யீட்டிற்குப் பிறகு அவற்றைப் பயன்படுத்தி பலவிதமான மொழிகளுக்குச சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் உருவாக்கப்பட்டன. அவை பின் வருமாறு:-

1.The MULTEXT 20 Project

2.The GRACE Project

3.The CRATER Project

EAGLES வழிகாட்டல்களால் அடக்கப்படாத மொழிகளின் சமீப காலத்திய சொல்வகைப்பாட்டு அடையாளக் குழுமங்களில் பின்வருவன அடங்கும்.

1.Tagset Design for Arabic.

2.Tagset Design for Chinese.

3.Tagset Design for Korean.

4.Tagset Design in the Paninijan Tradition.

6.6. தமிழ்த் தரவுத்தொகுதிகளை சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்தல்

6.6.1. தமிழ்ச் சொல்வகைப்பாடுகள்

ஒரு மொழியின் சொற்கள் சொல்லியல் அடிப்படையிலும், தொடரியல் அடிப்படையிலும், வகைப்பாடு செய்யப்படலாம். சொற்கள் எந்த இலக்கணக் கூறுக்காகத் திரிபுறும் வாக்கியத்தில் எங்கு, எவ்வாறு இடம் பெறுகின்றன. அவற்றின் செயற்பாடுகள் யாவை என்பதன் அடிப்படையில் அவை ஒரு சொல் வகுப்பைச் சார்ந்தன என நிர்ணயிக்கலாம். ஆங்கிலத்தில் “Part of speech tag” என்றும்,இதன் படி சொற்களை எட்டு வகுப்புகளாகப் பிரிப்பர்.

1. பெயர் - Noun
2. பெயரடை - Adjective
3. மாற்றுப்பெயர் - Pronoun
4. வினை - Verb
5. வினையடை - Adverb
6. முன்னுருபு - Proposition
7. இணைப்புக்கிளவி - Conjunction
8. வியப்புச்சொல் - Interjection

தற்காலத் தமிழ் இலக்கணத்தை ஆங்கில இலக்கண கண்ணோட்டத்தோடு நோக்கும் இலக்கண நூல்களில் ஆங்கிலச் சொல்வகைப்பாட்டின் தாக்கத்தினைக் காணலாம். பழந்தமிழுக்காக

எழுதப்பட்ட மரபிலக்கணங்களில் சொல் வகைகள் தற்காலத் தமிழுக்காக மாற்றியமைக்கப்பட வேண்டும் என மொழியியல் அறிஞர்கள் கருதுகின்றனர். இருப்பினும் தற்காலத் தமிழுக்கு எத்தனை சொல் வகுப்புகள் நிர்ணயிக்கப்பட வேண்டும் என்பதில் அவர்களிடையே ஒருமித்த கருத்து இல்லை.

மரபிலக்கணங்கள் தமிழ்ச் சொற்களை பெயர்ச்சொல், வினைச்சொல், இடைச்சொல், உரிச்சொல் என நான்கு வகுப்புகளாக கொண்டு, பெயரும்,வினையும் முக்கியமானதாகவும் இடையும்,உரியும் துணை வகுப்புகளாகவும் கொள்ளப்படுகின்றது. காலம் காட்டாமல் வேற்றுமை உருபுகளை ஏற்கும் சொற்கள் பெயர்ச் சொற்களாகவும், வேற்றுமை உருபுகளை ஏற்காது காலம் காட்டும் சொற்களை வினைச்சொற்களாகவும் கொண்டுள்ளனர். தனித்து இயங்கும் ஆற்றலின்றிப் பெயர்களையும்,வினைகளையும் இடமாகக் கொண்டு அவற்றின் முன்னும், பின்னும் வரும் வேற்றுமை உருபுகள், விகுதிகள், சுட்டெழுத்துகள், வினாவெழுத்துக்கள் ஆகியனவும் இடைச்சொல் எனப்பட்டன. பெயர்ச்சொல், வினைச்சொல் போல் தனித்து நின்று பொருளை உணர்த்தாமல் பெயர்ச்சொல்லையோ, வினைச்சொல்லையோ சார்ந்து நின்று அவற்றின் குணத்தை “மிகுதி” என உணர்த்த வல்லன உரிச்சொல்.

மரபிலக்கணத்தார் சொற்பகுப்பை ஆராய்ந்தால் சொற்களைப் பெயர்ச்சொல், வினைச்சொல் என்ற இரு பெரும் பிரிவுகளாகவும், அவற்றைச் சார்ந்து வரும் மூன்றாவது பிரிவுச்சொற்களை இடைச்சொல்,உரிச்சொல் என்று பிரித்தனர். மரபிலக்கணங்கள் உரிச்சொற்களைப் பெயர், வினை இவற்றிற்கு அடையாகக் கொள்கின்றனர். இவ்வாறு மரபிலக்கணங்கள் பெயர்ச்சொல், வினைச்சொல், இடைச்சொல், அடைச்சொல் என நான்கு வகைச் சொல் வகைகளை அடையாளம் காட்டுகின்றன. முற்காலத்தமிழை மரபிலக்கண வழி ஆய்கையில் இந்நான்கு வகை சொல் பகுப்புகள் போதுமானதாக இருந்தது.

தற்காலத் தமிழுக்கான சொல்வகுப்புகள் பற்றி இராசேந்திரன் அவர்கள் தமது கட்டுரையில் விரிவாகப் பேசுகிறார். அவர்கருத்துகள் இங்குத் தொகுக்கப்பட்டுத் தரப்பட்டுகின்றன். தற்காலத் தமிழ் புதிய இலக்கணக் கூறுகளை தன்னகத்தே வளர்த்துக்

கொண்டுள்ளமையால் சொற்களை புதிதாக வகைப்படுத்த மொழியியல் அறிஞர்கள் வாதிட்டும், வகைப்படுத்தியும் உள்ளனர். [Asher:1982, லேமன்:1989, Lehman:1989]

லேமன் சொல் வகைகளைப் பற்றி விரிவாக ஆராய்கிறார். [Lehman,1989, 911] தற்காலத் தமிழ் மொழியில் எல்லா அகராதி அல்லது வேர் உருபுகளை நான்கு வகைகளாகப் பிரித்து அவைகளை 1.பெயர் வேர்கள் 2. வினை வேர்கள் 3. வினையடை வேர்கள் 4. பெயரடை வேர்கள் எனக்கொள்கிறார். இவற்றில் பெரும்பாலான வேர்கள் பெயர், வினை வேர்களில் அடங்கும். பெயர்கள் பெயர் வேர்களையும், வினைகள் வினை வேர்களையும் கொண்டிருக்கும். ஆனால் பின்னுருபுகள் பல வினையடைகள், அளவையடைகள், இணைப்புச் சொற்கள் என்பனவற்றை வடிவ அடிப்படையில் பெயர் அல்லது வினை வேர்களின் திரிபுற்ற அல்லது திரிபுறா வடிவங்களாகக் கொள்ளலாம். பழந்தமிழில் சொல் வகுப்புகள் குறைந்த காரணத்தினால் பல திரிபுற்ற அல்லது திரிபுறாத பெயர், வினை மற்றும் பெயரடை வேர்கள் தற்காலத் தமிழில் பின்னுருபுகள், வினையடைகள், பெயரடைகள், அளவையடைகள் போன்ற பல சொல் வகுப்புகளாக மீள்ளாய்வு செய்யப்பெற்றன. நாம் அறிந்தபடி தற்காலத் தமிழில் இரண்டு சொல் வகுப்புகள் தான் திரிபுகளை ஏற்கவல்லது. அவை பெயரும் வினையும் ஆகும். பெயர்கள் வேற்றுமையுருபு, எண் இவற்றிறகாகத் திரிபுறும். வினைகள் காலம், எண், பால், இடம் மற்றும் சிலவற்றிற்காகவும் திரிபுறும். சொல் வகுப்புகளில் பெயர்கள் சொல்லாக்கத்தை காட்டுகின்றன. பெயர்களை வினைகளிலிருந்து பெறலாம்.

ஆஷர் [1982:101,102] ‘Operational definitions for word classes’ என்ற தலைப்பில் தமிழ்ச் சொற்களை ஆறு வகையாகப் பிரிக்கிறார்: 1. பெயர்ச்சொல் (Noun), 2. மாற்றுப்பெயர் (Pronoun), 3. வினைச்சொல் (Verb), 4. பெயரடை (Adjective), 5. பின்னுருபு (Post position), எண், அளவையடை, 6. இடைச்சொற்கள் (Particles)

இதையே கோதண்டராமன் (1989) தனியே இயங்க வல்ல சொற்களைத தலைமை இலக்கணக் கூறுகளாகக் கொண்டு பத்து வகுப்புகளாகப் பிரிக்கிறார். அவையாவன: 1. பெயர், 2. வினை, 3. பெயரடை, 4. வினையடை, 5. வல்லடை, 6. இணைப்பான், 7. உணர்ச்சி வெடிப்புச்சொல், 8. உணர்வு குறிப்புச்சொல், 9. விளிப்புச்சொல், 10. விளி ஏற்புச்சொல்

மேலும் “விகுதி”, “ஒட்டு” முதலானச் சார்புநிலைக் கிளவிகளைத் துணைமை இலக்கணக் கூறுகளாகக் கொண்டு அவற்றின் “வடிவுக்கும் இயல்புக்கும் ஏற்ப” 1. விகுதி (Suffix), 2. சொல்லுருபு (postposition), 3. வினைசார் கிளவி (Verb Participle), 4. ஒட்டு (Clitic), 5. சாரியை அல்லது நிரப்பி (Fillers) என ஐந்தாக வகைப் படுத்துகிறார்.

லேமன்(1989:9) உருபனியல் மற்றும் தொடரியல் இலக்கண இயல்புகள் அடிப்படையில் தமிழ்ச்சொற்களை 1. பெயர்கள் (Nouns), 2. வினைகள் (Verbs), 3. பெயரடைகள் (Adjectives), 4. வினையடைகள் (Adverbs), 5. பின்னுருபுகள் (Post position), 6. அளவையடைகள் (qantifiers), 7. வரையறு அடைகள் (Determiners), இணைப்புக் கிளவிகள் (Conjunctions) என எட்டாக சொல் வகுப்புகளைப் பிரிக்கிறார்.

இவ்மூவறிஞர்களும் பெயர், வினை, பெயரடை என்ற சொல் வகுப்புகளைத் தேர்ந்தெடுப்பதில் உடன்படுகிறார்கள். கோதண்டராமனும், லேமனும் வினையடையைச் சொல்வகுப்பாகக் கொண்டு ஆக்ஷரிடமிருந்து வேறுபடுகின்றனர். இணைப்பான் என்று கோதண்டராமனும், இணைப்புக்கிளவி என்று லேமனும் குறிப்பிடும் சொல்வகுப்பை ஆஷர் சொல்வகுப்பாகக் கொள்ளவில்லை. ஆஷரும், லேமனும் பின்னுருபுகளைச் சொல்வகுப்பாகக் கொள்ள கோதண்டராமன் பின்னுருபுகளைச் சொல்லுருபுகள் எனக்குறிப்பிட்டு சார்நிலைக்கிளவியாக வகைப்படுத்துகிறார்.

பெயர்ச்சொல்

ஆஷர் பெயர்ச்சொல் பிரிவை உருபனியல் மற்றும் தொடரியல் அடிப்படையில் சொல் வகுப்பாக நிறுவுகிறார். பெயர்ச்சொற்கள் உருபனியல் அடிப்படையில் வேற்றுமை ஒட்டுகளையும், பன்மை ஒட்டு “கள்” என்பதையும் ஏற்கும். தொடரியல் அடிப்படையில் பெயர்கள் பின்னுருபுத்தொடரின் தலைச் சொல்லாகவும், ஒரு வாக்கியத்தின் எழுவாயாகவோ, செயப்படு பொருளாகவோ செயல்படும். எழுவாயாக வரும் பெயர்ச்சொல்லின் இடம், எண், பாலுக்கு தகுந்தாற்போல் வினையின் இடம், எண், பால், ஒட்டின் தேர்வும் அமையும். வேற்றுமை உருபுகளையும் ஆக,ஆன கிளவிகளை ஏற்பனவும், எழுவாயாகவும், பயனிலையாகவும் செயல்பட வல்லன பெயர்கள் என்று கோதண்டராமன் குறிப்பிடுகிறார். லேமன் வேற்றுமை

பின்னொட்டுகளையும் ஆக,ஆய் பின்னொட்டுகளையும் ஏற்பன பெயர்கள் என்று விளக்குகிறார். மேலும் குறை உருபனில் உள்ள பல பெயர்கள் எல்லா வேற்றுமை பின்னொட்டுக்களையும் ஏற்காது என்பதையும் சுட்டுகிறார். பெயர்ச்சொல் வகுப்பில் மூவருக்கும் அடிப்படை பேதம் இருப்பதாகத் தெரியவில்லை. லேமன் மாற்றுப்பெயர், அளவுப்பெயர், எண்ணுப்பெயர் இவற்றை பெயர்ச்சொல்லின் துணைப்பிரிவாகக் கொள்கிறார்.

மாற்றுப்பெயர்

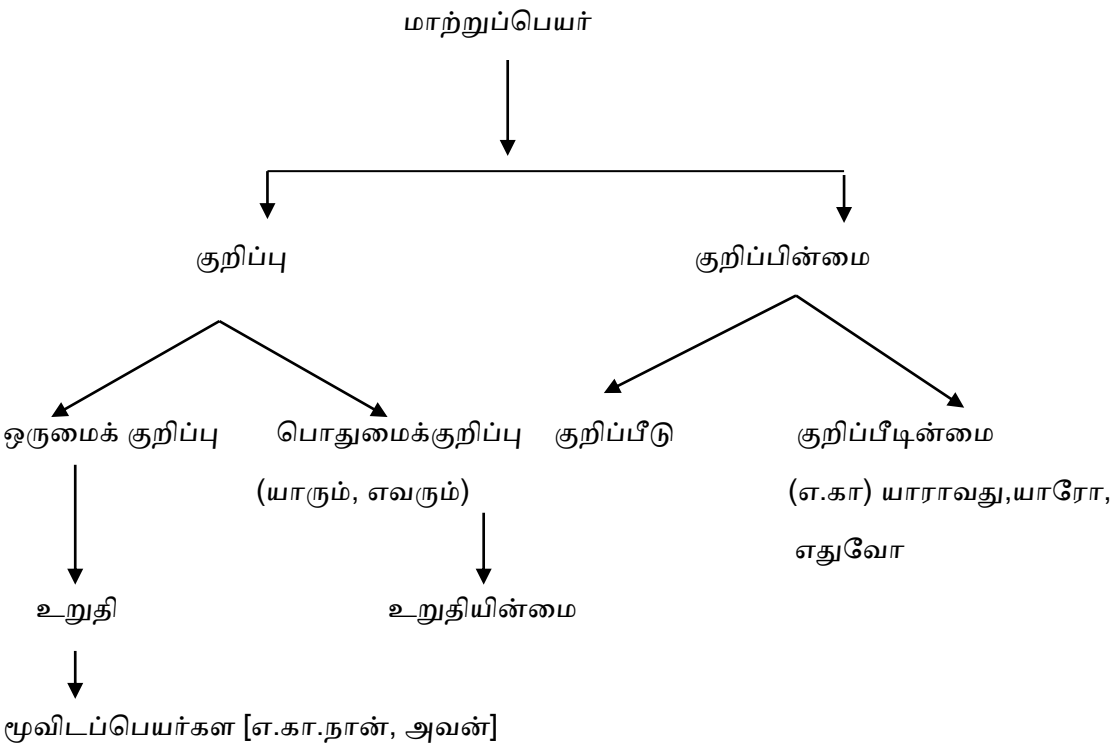
பெயர்ச்சொல்லுக்குப் பதிலாக அல்லது மாறாக வரும் சொல் மாற்றுப்பெயர் எனப்படும். இப்பெயர் முக்கியமாகப் பொருட்குறிப்பு (Signification) அடிப்படையில் பெயரிலிருந்து வேறுப்படும். ஒன்றையோ, பெயர்ச்சொல்லையோ குறிக்கும். ஆனால் மாற்றுப்பெயருக்குத் தக்கவாறு வெவ்வேறு ஆட்களையோ, பொருளையோ குறிக்கும். இதன் காரணமாய் ஒரு மொழியின் பெயர்ச்சொல்லின் எண்ணிக்கை ஏராளமாயிருந்தாலும், மாற்றுப்பெயர்கள் குறைவாகவே உள்ளது.

ஆஷர் மாற்றுப்பெயர்களைத் தனிச்சொல் வகுப்பாகக் கையாள்கிறார். பெயர் சொல்லின் வரையறைகளில் பல இயல்புகளைக் கொண்ட ஒரு நெருங்கிய குழுமத்தைச் சார்ந்திருப்பவை மாற்றுப்பெயர்கள். பெயர்ச்சொற்கள் ஏற்கும் அதே வேற்றுமைப் பின்னொட்டுகளை அவற்றுடன் சேர்க்கலாம். ஆவை பின்னுருபுத் தொடரின் தலைச் சொல்லாக அல்லது எழுவாயகவும், செயப்படுப்பொருளாகவும் செயற்படலாம். ஒரு வாக்கிய அமைப்பில் வினையின் இடம், எண், பால் ஒட்டைத் தேர்வு செய்யும்.

லேமன் பெயர்த் தொடரில் அடை எடுக்காது தனித்து வரும் பெயர்களை மாற்றுப் பெயர்களாகக் கொள்கிறார். மாற்றுப்பெயர்களைத் தனிநிலை மாற்றுப்பெயர்கள், [எ.கா.நான், அவன்] , ஆக்க மாற்றுப்பெயர்கள் [எ.கா.யாரோ, யாரும்] என்றும் வகைச் செய்கிறார். மூவிடப்பெயர்கள் [Personal Pronoun], சுட்டுப்பெயர்கள் [Demonstrative Pronouns], வினாப்பெயர்கள் [interrogative Pronouns], தன்வயப்பெயர்கள் [Reflexive Pronoun] என பலவகையாகப் பிரிக்கப்படுகிறது. மூலவகையான மாற்றுப்பெயர்களைக் குறிப்பு (Referential),

உறுதி (Definiteness), குறிப்பீடு (Specificity) என்ற பொருண்மைக் கருத்துகள் அடிப்படையில் வனைப்பாடு செய்யலாம்.

லேமன் ,1989:93



மூவிடப்பெயர்கள் காலத்தோடும், இடத்தோடும் தொடர்பு கொண்டிருக்கும். பேசுபவசைத் தன்மை இடம் என்றும், கேட்பவரை முன்னிலை இடம் என்றும், பேசப்படுபவரைப் படர்க்கை இடம் என்றும் செய்யலாம். சுப்ல கூடுதலான பாகுபாடுகளைக் கொண்டு படர்க்கையிடம், தன்மை, முன்னிலை இடங்களிலிருந்து வேறுபடுகின்றது. தன்மை, முன்னிலை, படர்க்கை ஆகிய மூன்றும் அடிப்படையில் ஒருமை, பன்மை என வேறுபடுகின்றது. முன்னிலை இடங்களும், படர்க்கை இடங்களும் தகுதி அடிப்படையில் உயர்வு, உயர்வின்மை என வேறுபடும். படர்க்ளை

இடப்பெயர்கள் மட்டுமே பால் அடிப்படையில் ஆண், பெண் என்றும், சுட்டு அடிப்படையில் அண்மை, சேய்மை என வேறுபடும்.

இதற்கான அமைப்பை கீழே அட்டவணை மூலம் காணலாம்

இடம்	ஒருமை	பன்மை
தன்மை	நான்	நாம்,நாங்கள்
முன்னிலை	நீ, நீர்	நீங்கள்
படர்க்கை	அவன்,இவன்,அவள்,இவள், அவர்,இவர்,அது,இது	அவர்கள்,இவர்கள் /அவை,அவைகள் இவை,இவைகள்

தான், நாம், தாங்கள் என்ற தன்வயப்பெயர்கள் நான்காவது இடத்தைச் சாரும். எழுவாயால் வரும் படர்க்கைச் சொல்லுக்கு மாறாய் அல்லது பதிலாய் தன்வயப்பெயர்கள் வரும்.

வினாப்பெயர்களுக்குத் தீர்மானிக்கப்பட்ட குறிப்பு இல்லை. வினாப்பெயர்களைக் குறிப்பீடு பெற்ற (Specified) வினாப்பெயர். குறிப்பீடு பெறாத (non specified) வினாப்பெயர் எனப்பிரிக்கலாம். அவை உயர்திணை, அஃறிணை வேறுபாட்டை வடிவத்தில் காட்டுகின்றன. (எ.கா. *யார், என்ன*) குறிப்பீடுப்பெற்ற வினாப்பெயர்கள் படர்க்கையிடம் எண், பால் இவற்றின் வேறுபாட்டை வடிவத்தில் காட்டுகின்றன. (எ.கா. எவன், எவள், எவர், எது, எவை) உம் என்ற குறைச்சொல் சேர்ந்த வினாப்பெயர்களுக்கு எடுத்துக்காட்டுகளாகும். குறைச்சொல் ஒ-சேர்ந்த வினாப்பெயர்களான யாரோ, ஏதோ, என்னவோ என்பன குறிப்பீடு பெறாத உறுதியற்ற மாற்றுப்பெயர்களுக்கு எடுத்துக்காட்டுகளாய் அமையும்.

அளவுப்பெயர்

சில, பல, எல்லாம், எல்லோரும் என்ற பெயர்கள் அளவுப்பெயர்கள் பிரிவைச் சாரும். இவை பெயர் அடைகளாகச் செயல்பட வல்லன (எ.கா. சில மனிதர்கள்). பெயர்களுக்குப்

பின்னால் வந்து வேற்றுமை ஒட்டுகள் எடுக்க வல்லன (எ.கா. பேனாக்கள் சிலவற்றை வாங்கினேன்). சில, பல என்பவைகளைப் பெயரடைகளாகக் கொண்டு முறையே சிலவை, பலவை என்பனவற்றை மாற்றுப் பெயரொட்டாக்கம் பெற்ற பெயர்களாக ஆயலாம் (எ.கா. சில+அவை=சிலவை)

எண்ணுப்பெயர்

எண்ணுப்பெயர்கள் ஆதார எண் (Cardinal Number), முறைமை எண் (Ordinal number) என இருவனைப்படும். முறைமை எண் ஆதார எண்ணிலிருந்து ஆவது அல்லது ஆம் என்ற குறைச்சொல்லின் (Clitics)சேர்க்கையால் ஆக்கப்படும். ஆதார எண்களிலிருந்து மாற்றுப் பெயரொட்டாக்கம் பெற்ற எண்ணுப்பெயர்களை (Pronominalzed Cardinal Number) பெறலாம். (எ.கா. ஒருவன், ஒருத்தி, ஒருவர்)

வினை

வினையைச் சொல்வகுப்பாகக் கொள்வதில் மொழியியல் அறிஞர்களிடையே கருத்து வேறுபாடு இல்லை. வினைகள் காலப்பின்னொட்டுகளையும், இடம் - எண் - பால் பின்னொட்டுகளையும் பிற பின்னொட்டுகளையும் எடுக்க வல்லன் சில வினைகள் குறை உருபனியலை உடையன. அதாவது எல்லா வினைப்பின்னொட்டுகளையும் அவை ஏற்பதில்லை. இணைப்பு வாக்கியங்கள் (Copular Sentences) (எ.கா. அவன் மாணவன்) தவிர்த்த பிற வாக்கியங்களில் கட்டாயமாக வரும் பகுதி வினையாகும். வினைகளை உருபனியல், பொருண்மையியல் இவற்றின் அடிப்படையில் வெவ்வேறு வகையாய் பாகுபாடு செய்யலாம். காலம் மற்றும் பிறவினைப் பின்னொட்டுகளுடன் சேரும்போது ஏற்படும் உருப ஒலியனியல் மாற்றங்கள் அடிப்படையில் வினைகளை மெல்வினை, இடைவினை, வல்வினை என்ற மூன்று வகைகளாகவும் அதற்கு மேற்பட்ட வனைகளாகவும் பிரிக்கலாம். வடிவ மற்றும் செயற்பாடு அடிப்படையில் வினையை முற்றுவினை (finite verb) (எ.கா. வந்தான்) மற்றும் எச்சவினை (non-finite berb) (எ.கா. வந்த, வந்து) என்று பிரிக்கலாம். எச்சவினை பெயரின் அடையாக வருமா அல்லது வினையின் அடையாக வருமா என்பதன் அடிப்படையில் பெயரெச்சம் (Adjectival or relative participle form) (எ.கா. வந்த பையன்) என்றும், வினையெச்சம் (adverbial or verbal

participle form) (எ.கா.வந்து போனான்) என்றும் வகைப்படுத்தலாம். தன்வினை (எ.கா.ஆடு) பிறவினை (எ.கா. ஆட்டு) என்ற பாகுபாடு பொருண்மையியல் அடிப்படையில் அமைந்தது. வினையினை தான் செய்யின் தன்வினை என்றும், பிறர் செய்யின் அது பிறவினை என்றும் விளக்கப்பட்ட பொருண்மை வேறுபாடு அவ்வளவு பொருத்தமானதல்ல. (பரமசிவம்,1983:2-3) அடிப்படையிலான வினைகளைத் தனவினைகள் என்றும், (எ.கா. ஓடு, காண்) அவற்றினின்றும் ஒட்டாக்கத்தால் ஆக்கப்பெற்ற வினைகளை (எ.கா. ஓட்டு, காட்டு) பிற வினைகள் என்றும் கொள்ளலாம். வாக்கியத்தில் பயன்படுத்துகையில் செயப்படுபொருளை எடுக்குமா? என்பதன் அடிப்படையில் வினைகள் செயப்படுபொருள் குன்றியவினை (எ.கா. போ, வா), செயப்படுபொருள் குன்றாவினை (எ.கா. படி, அடி) எனப்பாகுப்படுத்தலாம். பெயருக்கும் வினைக்கும் உள்ள காரக உறவுகளை (Case Relations)கொண்டும், இணைதிறன் (Valency))கொண்டும் வினைகளை வகைப்பாடு செய்யலாம்.

பின்னுருபுகள்

ஆக்ஷர் பின்னுருகைச் சொல்வகுப்புகளில் ஒன்றாகக் கூறுகின்றார். வேற்றுமைப் பின்னொட்டுகளால் திரிபுற்ற பெயர்களுக்குப் பின்வந்து பின்னுருபுத் தொடராக வினையுடன் ஒரு செயற்பாட்டு உறவில் நிற்கும் என்பது சொல் பின்னுருபு என்று தொடரியல் விளக்கம் தரலாம். பின்னுருபுகள் இந்நிலையில் தனித்து நிற்கும் என்பதிலிருந்து கட்டுப்பட்டே நிற்கும் என்பது பலதரப்பட்ட அங்கங்கள் சேர்ந்த சொல்வகுப்பாகும். கோதண்டசாமன்(1989) சாஹ்நிலைக் கிளவி வகை என்ற துணைமைச்சொல் வகையின் கீழ் பின்னுருபைச் சொல்லுருபு எனக் குறிப்பிட்டு வகைபாடு செய்கிறார். பெயர்ச்சொல்லும், வினைச்சொல்லும் காலப்போக்கில் வேற்றுமைப் பின்னொட்டுகள் போல் சொல்லுருபுகளாய் பயன்படுத்தப்படுகின்றன. லேமனும்(1989) பின்னுருபுகளைத் தனிச்சொல் வகுப்பாகக் கொள்கிறார். திரிபுற்ற மற்றும் திரிபுறாத பெயாவடிவுகளும் வினையெச்ச வடிவுகளும் தான் பின்னுருபுகளாகும். எடுத்துக்காட்டாக, தமிழ் பல இடப்பொருள் செயற்பாடுகளைக் குறிப்பிட, இடத்தைக் குறிக்கும் பெயர்களைத் தான் பயன்படுத்துகின்றது. (எ.கா. இல், உள், முன், மேல், கீழ், அடியில்) ஒரு குறிப்பிட்ட சொல்லைப் பெயராகக் கருதுவதா, பின்னுருபாகக் கருதுவதா என்பதில் ஒருமித்த கருத்தில்லை. பின்னுருபுகளாக எடுத்தாளப்படும் பெயர்கள் எல்லாம் உருபனியல் அடிப்படையில் குறை

உள்ளவை. அதாவது அவை எல்லா வேற்றுமை உருபுகளையும் எடுப்பதில்லை. பல பின்னுருபுகளாய் வரும் பெயர் வடிவுகள் தொடரியல் அடிப்படையில் குறைப்பாடுள்ளவை. அதாவது அவை வாக்கியத்தில் பெயர் வரும் எல்லா இடங்களிலும் வருவதில்லை. எல்லாப் பெயர்களைப் போலவும் அவற்றின் இலக்கணச் செயற்பாடுகள் வருவதில்லை. எடுத்துக்காட்டாக, நடு, இடை என்ற பெயர்கள் 'மையம்' என்ற பொருளில் இல் என்ற இடவேற்றுமை பின்னொட்டு எடுத்து நடுவில், இடையில் என பின்னுருபுகளாக மட்டும் வருகின்றன. ஆனால் அவை எழுவாயாகவோ, செயப்படுபொருளாகவோ, பயனிலையாகவோ வருவதில்லை. இப்பெயர் வடிவங்களில் மிகப் பெரும்பான்மையானவையாய் வரும் நெருங்கிய பெயர் மற்றும் வினை வடிவங்களின் குழுமம் பெயர்த் தொடருக்கப் பின்வந்து பின்னுருபுத் தொடராய் அமையும். பின்னுருபாய் வரும் பல பெயர், வினை வடிவங்கள் முறையே தங்கள் பெயர், வினை தொடரியல் இயல்புகளை இழந்து விடுகின்றன. பின்னுருபுகளை அவற்றின் வடிவ மற்றும் அவை எந்தப் பெயர்த் திரிபுக்குப் பிறகு வருகின்றன என்பதில் அடிப்படையில் வனைப்பாடு செய்யலாம். எடுத்துக்காட்டாக, இல் ஒட்டு எடுத்து வரும் நடுவில்,இடையில் என்பன – உக்கு வேற்றுமைக்குப் பின்னர் வரும். (எ.கா. அவர்களுக்கு நடுவில்,இடையில்) வினைவழிப் பின்னுருபுகளை இறந்தகால வினை எச்ச வடிவில் வருவன. (எ.கா. இல்லாமல்,அல்லாமல்) காலங்காட்டா வினையெச்ச வடிவில் வருவன. (எ.கா. தவிர, போல, விட) என வகைப்படுத்தலாம்.

பெயரடைகள்

ஆக்ஷர், லேமன், கோதண்டராமன் ஆகிய மூவருமே பெயரடையை ஒரு சொல் வகுப்பாகக் கொள்கின்றனர். பெயரடையை ஒரு சொல் வகுப்பாகக் கொள்வதில் மொழியியல் அறிஞர்களிடையே ஒருமித்த கருத்து இல்லை. பெயரடைக்கு ஒரு செயல்முறை வரையறை செய்வது எளிதானதல்ல. பெயரடை தொடரியல் அடிப்படையிலான ஒரு சொல் வகையாகும். பெயரடை, பெயர்த்தொடரில் தலைப்பெயருக்கு முன்னதாக அதன் அடையாக வரும் (எ.கா. பெரிய மரம்), வரையறு அடை (Determiner) பெயரடைக்கு முன் வரலாம். (எ.கா. இந்தப் பெரிய வீடு) பயனிலை இடத்தில் பெயரடை வருவதானால் அது மாற்றுப்பெயர் ஒட்டாக்கம் பெற்று வர வேண்டும்.(எ.கா. அந்த மாளிகை பெரியது) சொல் வகுப்புகளில் பெயரடை குறைச்சொல் எடுப்பதில்லை. பெயரடைகளைத் தனிநிலைப் பெயரடை (எ.கா.நல்ல, பெரிய, கெட்ட, சிறிய),

கூட்டுநிலைப் பெயரடை மற்றும் ஆக்கப்பெயரடை (derived adjective) (எ.கா. அழகான, உயரமான) என வகைப்படுத்தலாம். தனிநிலைப் பெயரடைகள் மரபிலக்கணத்தில் குறிப்புப் பெயரெச்சங்களாகக் கொள்ளப்படுகின்றன. பெயருக்கு பின் வந்து அவற்றைப் பெயரடைகளாகச் செயலாற்றச் செய்யும் ஆன, உள்ள, அற்ற, இல்லாத (எ.கா.அழகான, அழகற்ற, அழகில்லாத) என்ற பின்னுருபுகளைப் பெயரடையாக்கப் பின்னொட்டுகளாகக் கொள்வதா? அல்லது பெயரெச்சங்களாகக் கொள்வதா? என்பது இலக்கணப் பிரச்சனை. பெயரடைகளும் (தெரிநிலை) பெயரெச்சத் தொடர்களும் செயற்பாடு அடிப்படையில் பெயருக்கு அடையாய் வருவன (எ.கா. அந்த நல்ல பையன், அந்த படித்தப் பையன்) ஆனால் (தெரிநிலை) பெயரெச்சங்கள் உடன், பின், பிறகு, போது, முன், மட்டும், வரைக்கும் முதலான உருபுகளுடன் வினையடைத் தொடர்புகளை உருவாக்கும். (எ.கா. வந்தவுடன், வந்தபின், வரும்முன், கூடியமட்டும்) பெயரடைகள் (குறிப்புப் பெயரெச்சங்கள்) இவ்வாறு செயல்படுவதில்லை. (பரமசிவம் (1983:194)) பரமசிவம் மரபிலக்கணங்கள் பெயரெச்சங்கள் எனக் குறிப்பிடும் தெரிநிலைப் பெயரெச்சம் (எ.கா. பாடின பெண்) குறிப்புப் பெயரெச்சம (எ.கா.நல்ல பெண்), எதிர்மறைப் பெயரெச்சம் (எ.கா. படிக்காத பெண்) என்ற மூன்றையும் ஆன ஒட்டாக்காப் பெயரடைகளையும் (எ.கா.அழகான பெண.) பெயரடைகளாகக் கொண்டு அவற்றில் தெரிநிலைப் பெயரெச்சத்தையும், எதிர்மறைப் பெயரெச்சத்தையும் தொடர் என்று குறிப்பிட்டு ஒரு வகையாகவும், குறிப்புப் பெயரெச்சத்தையும் ஆன ஒட்டாக்கப் பெயரடையையும் தனிச்சொல்லாகக் கொண்டு மற்றொரு வகையாகவும் பிரிக்கிறார்.

வினையடைகள்

ஆஷர் (Asher, 1982:101-102) ‘Operational Definition for Word Classes’ என்ற தலைப்பின் கீழ் சொல் வகுப்புகளில் ஒன்றாக வினையடையைத் தராவிடிலும் சொல்லாக்க உருபனியலைப் பற்றிப் பேசுகையில் அவர் (Asher, 1982:199-205) விடையடைகளின் ஆக்கம் பற்றிக் குறிப்பிடுகிறார். வாக்கியத்தில் அவை இடம் பற்றிக் கூறுகையில் இடவாக்கியங்கள் (Locative Sentences) மற்றும் உள்ளமை வாக்கியங்கள் (existential Sentences) அல்லாத பிற வாக்கியங்களில் வினையடை பொதுவாக எழுவாய் மற்றும் அயல் செயப்படுபொருளைத் தொடர்ந்தோ நேர்ச் செயப்படுபொருளுக்கு முன்னரோ வரும் என்று குறிப்பிடுகிறார். ஒரே

வாக்கியத்தில் வேறுபட்ட வினையடைகள் வந்தால் அவை வரும் நிரலைச் சரியாகக் கூறமுடியாது. பொதுவாக, கால வினையடைகள், இட வினையடைகளுக்கு முன்னர் வரும். இடவாக்கியங்கள் இவற்றில் வினையடை எழுவாய் முன்னர் வரும். கோதண்டராமனும், லேமனும் வினையடையைச் சொல் வகுப்பாகக் கொள்கின்றனர். வினையடைகளைத் தனிநிலை வினையடை (Simple adverbs) (எ.கா.நன்கு, உரைக்க) கூட்டுநிலை வினையடை அல்லது ஆக்க வினையடை (எ.கா. அழகாக, அழகாய்) எனப்பகுக்கலாம். திரிபுற்ற அல்லது திரிபுறாத பல பெயர்ச்சொல் மற்றும் வினைச்சொல் வடிவங்களை தொடரியல் அடிப்படையில் நெருங்கிய ஒரு குழு வினையடையாக மீள் ஆய்வு செய்யலாம். பின்னுருபுகளாக வரும் பல சொல் வடிவுகள் வினையடைகளாகவும் வரலாம். (எ.கா. முன்னர், முன்னால், பிறகு) திரிபுறாத ஆனால் குறைச்சொல் ஏ –ஐ விருப்பாக ஏற்றுவரும். பெயர்ச்சொல் வடிவுகள் (எ.கா. அப்பாலே, உள்ளே, அப்புறம்) இடவேற்றுமைக்காகத் திரிபுற்ற பெயர்ச்சொல் வடிவுகள் (எ.கா. பார்த்து, பிந்தி, முந்தி) என்பன வினையடைகளாகவும், பின்னுருபுகளாகவும் செயல்பட வல்லன. பெயரல்லாத வேறுசில சொல்வடிவுகளும் வினையடையாகச் செயல்பட வல்லன. எடுத்துக்காட்டாகச் சுட்டு மற்றும் வினாச் சுட்டுரபுகள் கொண்ட அப்படி, எப்படி, அப்போது, எப்போது, இங்கு, எங்கு, இன்று, என்று என்பனவற்றையும், பொதுவாக வினையடைகளாகக் கொள்ளலாம். சில இடங்களில் இவை வினையடையல்லாது பெயர்போல் செயற்பாடு செய்யலாம். (எ.கா. இன்று நல்ல நாள்) பல திரிபுற்ற மற்றும் திரிபுறாத பெயர் மற்றும் வினை வடிவங்கள் தொடரியல் அடிப்படையில் வாக்கியநிலை வினையடைகள் அல்லது வினையடை இணைப்பான்கள் என மீள் ஆய்வு செய்யப்படலாம். இவை வாக்கியத்தின் முன்னிடத்தில் வந்து இரண்டு வாக்கியங்களின் பொருண்மை உறவைக் கூறி நிற்கின்றன. எ.கா. அவன் நன்றாய்ப் படித்தான், ஆனால் வெற்றிபெறவில்லை. அப்படியும், ஆனால், இருந்தாலும், எதற்கும் என்பன வாக்கியநிலை வினையடைகளாகும்.

பெயர்ச்சொல்லுடன் ஆகு என்ற வினைச் சொல்லின் காலமிலா வினையெச்ச வடிவான ஆக அல்லது ஆய் சேர்ந்து கிடைக்கும் சொல் வடிவங்களை ஆக்ஷரும், கோதண்டராமனும் வினையடைகளாகக் கொள்கின்றனர். ஆனால் லேமன் ஆக,ஆய் என்பது வினையடையாக்கம்

செய்வது மட்டுமன்றி பிற செயற்பாடுகளிலும் வருவதைச் சுட்டிக்காட்டி ஆக,ஆய் என்பதை கட்டுநிலைப் பின்னுருபு அல்லது குறைச்சொல் என்று ஆயலாம் என்கிறார்.

பொருண்மை அடிப்படையில் வினையடைகளைப் பின்வருமாறு பகுக்கலாம்:

1. கால வினையடைகள் (எ.கா.இன்று, நாளை)
2. இட வினையடைகள் (எ.கா.இங்கு, அங்கு)
3. முறைமை வினையடைகள் (எ.கா.மெல்ல, நன்கு)
4. செயல்தொடர் வினையடைகள் (எ.கா.அடிக்கடி, மறுபடியும்)
5. அளவு வினையடைகள்

ஆனால் லேமனும், ஆஷ்ரும் அளவு வினையடைகளை ஒரு தனி வகுப்பாகக் கொள்கின்றனர். பரமசிவம்(1983) வினையெச்சங்களையும் ஒருவகை வனையடையாகக் கொள்கின்றார். வினையெச்சங்கள் காலக்குறிப்பினைக் கொண்டு விளங்குவதால் இவற்றை வினையடைகளாக எல்லோரும் கொள்வதில்லை.

அளவயடைகள்

ஆஷரும், லேமனும் அளவயடைகளைச் சொல் வகுப்பாகக் கொள்கின்றனர். அளவடைகளுக்கு உருபனியல் அடிப்படையில் ஒட்டு மொத்தமான வரையறை தர இயலாது. நெருங்கிய ஒரு குழுமச் சொல்வடிவுகள் (எ.கா. சற்று, முழு, கொஞ்சம், இத்தனை, அத்தனை, இவ்வளவு, அவ்வளவு, எவ்வளவு, நிரம்ப, நிறைய, மிகவும்) பெயர் அடையாக வர வல்லன. ஆனால் இவை வருமிடங்கள் ஒன்றானதல்ல. முழு என்ற அளவடை எப்பொழுதும் தலைப்புப் பெயருக்கு முன் அதனை அடுத்து வரும். (எ.கா.சற்று நேரம்), பெயருடன் ஆக,ஆன, இல்லாத, உடைய சேர்ந்த சொல் வடிவங்களின் முன்னர் மிகவும் வரும். (எ.கா. மிகவும் அழகாக, மிகவும் அழகான, மிகவும் அழகில்லாத) பிற அளவடைகள் பெயரடைூபெயர் என்ற தொடரின் முன்வரும். (எ.கா.கொஞ்சம்,பெரிய,பாத்திரம்) முழு, இத்தனை, அத்தனை, எத்தனை, மிகவும், தவிர்த்த பிற அளவடைகள் வினையின் அடையாக வினை முன்னிடத்தில் வரும். எ.கா. சற்று மறந்தான், நிறைய சாப்பிட்டான்). மிகவும் என்பன பெயரடையின் அடையாக வந்து பெயரடைத் தொடரில் காணப்படும். (எ.கா கொஞ்சம் சின்னக் கை) கோதண்டராமன் மிகவும் என்பதை பெயர், வினை, பெயரடை, வினையடை ஆகிய நான்கினுக்கு முன் அடையாக வரும். சொற்களை

வல்லடை என்ற தனிச்சொல் வகுப்பில் அடக்குகின்றார். (எ.கா. மிகவும் கறுப்பு, மிகவும் நல்ல, மிகவும் வேகமாக, மிகவும் பிடிக்கும்).

வரையறை அடைகள்

பெயருக்குப் பின்னர் வரும் இந்த, அந்த என்ற சுட்டுப்பெயரடைகள் சரையறை அடைகளாகும். ஆவை சுட்டப்படும் பொருள் அல்லது குறிப்புப்பொருள் (referent) பேசுபவர்களிடமிருந்து எவ்வளவு அண்மைப்பட்டு அல்லது சேய்மைபட்டு இருக்கின்றது என்பதைக் குறிக்கவல்லன. (எ.கா இந்தப் பையன், அந்தப் பெண்) கிரியாவின் தற்காலத் தமிழ் அகராதியில் இவை சுட்டுப் பெயரடைகள் என்று குறிப்பிடப்படுள்ளன.

இணைப்பான்கள்

கோதண்டராமனும், லேமனும் இணைப்பான்களைச் சொல்வகுப்பாகக் கொள்கின்றனர். "இரண்டு சொற்றொடர்களை அல்லது சொற்களை இணைப்பதாகவும் ஒரு முழுத்தொடரைத் தன்னோடு இணைப்பதாகவும் அமைந்த சொற்கள் இணைப்பான்கள்" எனக் கோதண்டராமன் விளக்குகின்றார். தமிழில் பொதுவாக குறைச்சொற்கள் (clitics) இருப்பினும் பலவினை வடிவங்களை தொடரியல் அடிப்படையில் இணைப்புச சொற்களாகக் கொள்ளலாம். (எ.கா. ஆனால், அல்லது, இல்லையென்றால்) இணைப்புச் சொற்கள் இருபெயரடைகளையோ, பெயர்களையோ இணைக்கப் பயன்படும். (எ.கா. பழைய ஆனால் நல்ல புத்தகம், ரேக்ஷ்மா இல்லையென்றால் கலா வருவாள். கிரியாவின் தற்காலத் தமிழகராதி இவற்றை இணைப்பு இடைச்சொற்கள் என்று குறிப்பிடுகின்றது.

குறைச்சொற்கள்

குறைச்சொற்கள் பற்றி ஆரோக்கியநாதன் (1982) விரிவாக ஆராய்ந்துள்ளார். கோதண்டராமன் குறைச்சொற்களைத துணைமைச்சொல் வகுப்பாகச் சார்புநிலைக்கிளவி வகையின் கீழ் **'ஒட்டு'** என்று குறிப்பிட்டு வகைப்படுத்துகிறார். அவர் "தொடரில் பல்வேறு இடங்களில் நின்று தொடர்ப் பொருளில் நிலைமாற்றம் செய்யும் மற்றும் பெயர் விகுதியாகவோ,

வினைவிகுதியாகவோ" கொள்ள இயலாத தான், உம், ஆ முதலான கிளவிகளை ஒட்டுகள் என்று குறிப்பிடுகின்றார். லேமன் குறைச் சொற்களைப் பற்றிப் பேசும் போது உருபனியல் அடிப்படையில் அல்லாமல் இலக்கணத்தின் ஒலியனியல் விதிகளின்படி சொறகளுள் ஒட்டப்பெறும் கட்டு வடிவங்கள் குறைச்சொற்கள் எனக் குறிப்பிடுகின்றனர். அவை திரிபு அல்லது ஆக்கக் கூறுகளின் பிரதிபலிப்பல்ல. திரிபு அல்லது ஆக்கப் பின்னொட்டுகளைப் போல் அவை ஒரு குறிப்பிட்ட சொல் வகுப்பை மட்டும் சார்ந்து வருவன அல்ல. தமிழில் எல்லாக் குறைச்சொற்களும் பின் குறைச் சொற்களாகும். பெயரடை அல்லது பெயரின் அடையாய் செயல்படும் பெயர்கள் தவிர பிற தொடரியல் வகைப்பாடுகளின் தலைமைச் சொற்களுடன் குறைச் சொற்கள் ஒட்டப்பெறலாம். ஒரு குறிப்பிட்ட ஒலியனியல் வடிவம் கொண்ட குறைச்சொல்லுக்குப் பல பொருண்மைச் செயல்பாடுகள் இருக்கும். உம், ஓ, ஏ, தான், ஆ என்பன குறைச்சொல் வகுப்பின் பாற்படும். ஆக்ஷர் ஏ, தான் ஆகிய தேற்றச் குறிகளையும் (emphatic markers) ஆக என்ற வினாக் குறியையும் (interrogative marker) உம், ஓ ஆகிய இணைப்பான்களையும் பலதரப்பட்ட முக்கியமான வாக்கிய உறுப்புகளுடன் ஒட்டத் தகுந்த மாற்றம் கொள்ளாத கட்டு வடிவங்கள் எனக் குறிப்பிட்டு அவற்றை இடைச்சொற்கள் (Particles) என்ற சொல் வகுப்பில் அடக்குகின்றார். கிரியாவின் தற்காலத் தமிழகராதி இவற்றை இடைச்சொற்கள் எனக் குறிப்பிடுகின்றது.

வினைசார் கிளவி

கோதண்டராமன் வந்தபின், வந்தபோது, வந்த உடன் ஆகியவற்றில் காணும் பின், போது, உடன் ஆகிய கிளவிகளை வினைசார் கிளவி என துணைச்சொல் வகையாகச் சார்புநிலைக் கிளவி வகையின் கீழ் தருகின்றார். "பெயர்த்தன்மை திரிந்து பெயரெச்சத்தின் பின் விகுதி போல நின்று வினை எச்சங்களை அமைக்கும் கிளவிகளும், வினைசார் கிளவிகள்" என்று குறிப்பிடுகின்றார். செய்யத்தக்க, செய்யக்கூடிய, செய்ய வேண்டிய முதலான கூட்டுநிலைப் பெயரெச்சஆகளில் காணும் தக்க, கூடிய, வேண்டிய என்பனவற்றையும் வினைசார் கிளவிகளாகக் கொள்ளத்தகும் என்கிறார். கிரியாவின் தற்காலத் தமிழகராதி இயற்றையும் இடைச்சொல் என்றே குறிப்பிடுகின்றது. வேற்றுமைப் பின்னொட்டுகளையும் (எ.கா. ஐ, ஆல், கு) அன், ஆன், அள், ஆள்

முதலான பால் இட ஒட்டுகளையும் ஆன, ஆக என வரும் பெயரடை, வினையடை விகுதிகளையும் இந்த வகுப்பில் அடக்குகின்றார்.

சாரியை அல்லது நிரப்பி

"தமக்கென்று அகராதிப் பொருளோ, இலக்கணப் பொருளோ இல்லாதனவாய் சொற்கள் ஒன்றோடொன்று சேரும்போது சொற்சேர்க்கைக்கு இசைவாக நின்று சொல் அமைப்புக்கு உதவுவனவாய் அமைந்த கிளவிகள் சாரியை அல்லது நிரப்பி என்று பெயர்பெறும்" என்று கோதண்டராமன் இச்சொல் வகுப்பை வரையறை விளக்கம் செய்கிறார். வீட்டடினை, ஊரினை என்பனவற்றிலுள்ள இனை என்பதையும், வந்தனன், சென்றனன என்பனவற்றிலுள்ள அன் என்பதையும் சாரியை கிளவிக்கு எடுத்துக்காட்டுகளாகக் காட்டுகின்றார்.

மேலே கூறப்பட்ட அனைத்துச் சொல் வகைப்பாடு தமிழுக்குத் தேவையான அடிப்படைச் சொல் வகைப்பாடுகளாகும். இதை தவிர பல நுண்மையான வகைப்பாடுகளையும் உட்படுத்தலாம். எடுத்துக்காட்டாக, வினையை, வினையெச்சம், பெயரெச்சம், வினைமுற்று, துணைவினை என மேலும் வகைப்படுத்தலாம். இச்சொல் வகைப்பாடுகள் சொல்வகை அடையாளக் குழுமத் தேர்வுக்கு அடிப்படையாய் அமையும்.

6.6.2 தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

6.6.2.1 CIIL நிறுவனத்தின் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

இந்திய மொழிகளுக்கு தரவுத்தொகுதிகள் உருவாக்குவதில் முன்னோடியாக விளங்கிய இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்திய மொழிகளுக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கித் தரவுத்தொகுதிகளை அடையாளப் படுத்தியுள்ளது. 2006-இல் இந்தி தரவுத்தொகுதிக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமமாக முடிவு செய்யப்பட்ட சொல் வகைப்பாட்டு குழுமத்தின் பட்டியல் கீழே தரப்பட்டுள்ளது.

CIIL Hindi Corpora Tagset Finalized in August 29.2006

NOUN

1. NN Commom Noun
2. NNP Proper Noun (Palce, Person,etc.…)
3. NC Noun Compound
4. NAB Noun Abstgract
5. CRD Cardinal number
6. ORD Ordinal number

PRONOUN

7. PRP Personal Pronoun
8. PRI Indefinite Pronoun
9. PRR Pronoun Reflexive
10. PRL Relative Pronoun
11. PDP Demonstrative Pronoun

VERB

12. VF Verb Finite Main
13. VNF Verb non-finite Adjectival/ Verb non-finite Adverbial
14. VAX Verb Auxiliary
15. VNN Gerund/Verb non-finite nominal
16. VINF Verb Infinitive
17. VCC Verb Causative
18. VCD Verb Double Causative

ADJECTIVE

19. ADD Adjective - declinable
20. ADI Adjective - Indeclinable

ADVERB

21. ADV Adverb
22. ADL Adverb of Location

OTHERS

23. QWO Question Words
24. PPO Post position
25. INT Intensifier
26. QTF Quantifier
27. CNJ Conjunction
28. INJ Interjection
27. IND Indiclinable
30. QOT Quotative
31. PRT Particle
32. NEG Negative
33. RDP Reduplication
34. FWD Loan Word [non – nativized words]
35. IDM Idiom
36. PRO Proverb

6.6.2.2 AU-KBC நிறுவனத்தின் தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமம்

AU-KBC நிறுவனம் தமிழுக்கான ஒரு சொல் வகைப்பாட்டுக் குழுமத்தை உருவாக்கியுள்ளது. இது எல்லா இலக்கண மற்றும் சொல்சார் உறுப்புகளையும் உள்ளடக்கிய குழுமம் ஆகும். இது தமிழ்ப்பல்கலைக் கழக மொழியியல் வல்லுநர்கள் உதவியால் AU-KBC -இன் NLP குழுவினரால் 2001-இல் திட்டமிடப்பட்டு உருவாக்கப்பட்டதாகும். பின்வரும் பட்டியல் தமிழின் சொல் வகைப்பாட்டு அடையாளக் குழுமமாகும்.

TAMIL TAGSET

No	Description	TAG	Examples	Suffixes
1	Noun Singular	N.Sg	மரம் (tree)	
2	Noun Plural	N.pl	மரங்கள் (trees) பூக்கள் (flowers)	

3	Pronoun	Pr.Sg	avanY (he), nAnY(I)	
4	Pronoun Oblique from	Pr.Obl.Sg	என்(mine)	
5	Pronoun plural	Pro.Pl	அவர்கள் (they)	
6	Pronoun Plural Oblique	Pro.Obl.Pl	எங்கள் (ours), உங்கள் (yours)	
7	Euphonic Increment	Eu	பசுவின் (of the cow)	
8	Acusative	Acc	அவனை (him accusative)	ஐ
9	Dative	Dat	அவனுக்கு (to)	உக்கு, கு, க்கு
10	Instumental	Inst	சாவியால் (by the key)	ஆல்
11	Sociative	Soc	அவனோடு (with him), அவனுடன் (with him)	ஓடு, உடன்
12	Locative	Loc	கையில் (in the story)	இல்
13	Ablative	Abl	ஊரிலிருந்து (from city/town	ல்லிருந்து
14	Benefactive	Ben	அவளுக்காக (for her)	உக்காக
15	Genetive	Gen	பசுவின் (of the cow), உடைய (of), அவனது (his)	இன், உடைய, அது
16	Vocative	Voc		அட, இவளே
17	Clitics	Cl		ஆ,ஏ,ஓ,உம்,தான்
18	Verb	V	ஓடு (run)	
19	#Transitive Verb	VT	திருப்பு (turn sth)	

20	#Intransitive Verb	VI	திரும்பு (turn)	
21	Causative	VC	தெரிவி (inform)	
22	Infinitive	Inf	செய்ய (todo)	அ
23	Singular Negative [imperative]	Neg	செய்யாதே (you don't do)	ஆதே
24	Plural Imperative	V.pl	செய்யுங்கள் (you do-honor plu)	உங்கள்
25	Plural Negative [imperative]	Pl..Neg	செய்யாதிருங்கள் (you do-hon or plu)	ஆதிருங்கள்
26	Imperative+Suffixes	Imp	வாட, அம்மா	
27	Present Tense	Pre		கிற், க்கிற், கின்ற், க்கின்ற்
28	Past Tense	Pst		த், ந்த், இன்,த்த், இ, ட், ற்
29	Future Tense	Fut		வ்,ப், ப்ப், உம்
30	First person Singular	1s		என்
31	First person Plural	1pl		ஓம்
32	Second person Singular	2s		ஆய்
33	Second person Plural	2pl		ஈர், ஈர்கள்
34	Third person masculine Singular	3Ms		ஆண்
35	Third person feminine Singular	3 Fs		ஆள்
36	Third person Honorific Singular	3 Hs		ஆர், அர், அனர்

37	Third person Honorific Plural	3 H pl		ஆர்கள்
38	Third person singular Neuter (non-human)	3 N S		awu, wu, um
39	Third person plural Neuter (non-human)	3 Npl		அன, உம்
40	Third person Neuter Future Negative	3NFut Neg	விடாது (it/they), (won't)	ஆ (ஆது?)
41	Future Negative human	Hum Neg	மாட்டேன் (I don't)	மாட்ட்
42	Present/Past Negative (non-future)	Nfut Neg	வரவில்லை (didn't come)	இல்லை
43	#Optative	Opt	வாழ்க (long live)	க, க்க
44	Verbal Participle	Vbp	செய்து (done)	து, ந்து, etc.,
45	Verbal Participle Negative	Vbp Neg	செய்யாமல் (non-done)	ஆமல், ஆது
46	Contitional Positive	Cnd	செய்தால் (if done)	ஆல்
47	Conditional Negative	Cnd Neg	செய்யாவிட்டால் (if not done)	ஆவிட்டால்
48	Adjectval Participle Present/Past	AdjP	செய்த (do pre/pst+sth), செய்கிற (do pre/Pst-th)	த, கிற
49	Adjectival Participle Future	Adjp Fut	ceVyyum (do Future sth)	Um
50	Adjectival Participle Negative	Adjp Neg	செய்யாத (don't do sth)	ஆத
51	Verbal Noun Untensed	Vbn Unt	செய்யல் (doing),	அல், தல், த்தல்,

			செய்தல் (doing), படிக்கல் (reading), படித்தல்l (reading), செய்கை (doing), படிக்கை (reading)	க்கல், கை, க்கை
52	Verbal Noun Tensed Positive	Vbn	செய்தது (have done)	ஆது
53	Verbal Noun Tensed Negative	Vbn Neg	செய்யாத்து (have not done)	ஆதது
54	Participal Noun	PpIN	செய்தவன் (one who did)	அன், அள்,து, அர்,அர்கள்.ஐ
55	Adjctival Noun	AdjN	நல்லவன் (good man)	அன், அள்,து, அர், அர்கள்.ஐ
56	Selective (Concessive?)	Sel	அவன், ஆவது(at least him)	ஆவது
57	Ordinal Suffix	Ord	முதலாவது(First)	ஆவது
58	Reportative Suffix	Rep	-	ஆம்
59	Complementizer	Cmp	-	என்று, என்பது, ஆக
60	Auxiliary	Aux	-	-
61	Conjunction	Conj	-	ஆனால், இல்லை என்றால், அல்லது
62	Adjective	Adj	அழகான(beautiful with adj suffix)	ஆன
63	Adverbs	Adv	அழகாக(beautiful	ஆக

			with adv Suffix)	
64	Postposition	PSP	மேல் (above), கீழ் (below)	
65	Quantifiers	Q	ஒன்று (one), சில (few), கொஞ்சம் (some),	
66	Determiners	Det	அந்த (that), இந்த (this)	
67	Wh. Words	INT	என்ன(what), எவை(which), எப்படி(how)	
68	Unknown	UNK	-	-

6.6.2.3 அமிர்தா பல்கலைக்கழகத் தமிழுக்கான சொல்வகை அடையாளக் குழுமம்

அமிர்தா தொழில் நுட்ப கல்வி நிறுவனத்தில் தமிழை மொழிபெயர்ப்பதற்காக நுட்பங்களைப் பயன்படுத்தி அமிர்தா தொழில் நுட்ப கல்லூரி பேராசிரியர்களான தனலெட்சுமி, அனந்த் குமார், சோமன் மற்றும் தமிழ்பல்கலைக்கழக மொழியியல் துறைத்தலைவர் ச. இராஜேந்திரன் போன்றோர்களின் கூட்டு முயற்சியால் உருவாக்கப்பட்ட சொல்வகை அடையாளப்படுத்திகளின் அட்டவணையைக் கீழேக் காணலாம்.

S.No	POS	DESCRIPTION
1	NN	NOUN
2	NNC	COMPOUND NOUN
3	NNP	PROPER NOUN
4	NNPC	COMPOUND PROPER NOUN
5	ORD	ORDINALS

6	CRD	CARDINALS
7	PRP	PRONOUN
8	PRIN	PRONOUNINTROGATIVE
9	PRID	PRONOUN INDIFINITE
10	ADJ	ADJECTIVES
11	ADV	ADVERB
12	VNAJ	VERB NON-FINITE ADJECTIVES
13	VNAV	VERB NON-FINITE ADVERBS
14	VBG	VERBIAL GERLND
15	VF	VERB FINITE
16	VAX	VERB AUXILARY
17	VNIT	VERB IN FINITE
18	CNJ	CONJUNTION
19	CVB	CONDITIONAL VERB
20	QW	QUESTION WORD
21	COM	COMPLEMENTIZER
22	NNQ	QUANTITY NOUN
23	QTF	QUANTIFIERS
24	PPO	POSTPOSITION
25	DET	DETGERMINERS
26	INT	INTENSIFIER
27	ECH	ECHO WORDS
28	EMP	EMPHASTS
29	COMM	COMMA

30	DOT	DOT
31	QM	QUESTION MARKS
32	RDW	REDUPLICATION

இவ்வட்டவணையைக் கொண்டு உருவாக்கப்பட்ட சொல்வகை அடையாளப்படுத்தியைக் கொண்டு பல்வேறு அடிப்படையில் உருவாக்கப்பட்ட தொடர்களின் அமைப்பைக் கீழேக் காணலாம்.

பெயர்த்தொடர்

[அந்த <DET> (B-NP) அழகான <ADJ> (I-NP) பறவை <NN> (I-NP)] NP

பெயரடைத்தொடர்

[திரைப்படம் <NN> (B-AJP) சார்ந்த <ADJ> (I-AJP)] AJP

வினையடைத்தொடர்

[அருகே <ADV> (B-AVP)]AVP

இணைப்புச்சொல்

[ஆனால் <CNJ>(B-CJP)] CJP

[என்று<COM> (B-COMP)] COMP

வினைமுற்று தொடர்

[உள்ளது <VF> (B-VFP)] VFP

வினைமுற்றுப்பெறாத வாக்கியம்

[வெளி வந்த (VNAJ) (B-VNP)] VNP செய்திக் <NNC> < B-NP> குறிப்பு <I –NP> <NNC>

[விரைந்து <VNAV>(B-VNP)] VNP செய்தான் <VF>

gerundial chunk

தொழிற்சாலை <NN> [அமைப்பதில் <VBG>(B-VGP)] VGP தாமதம் <NN>

ஒரு விரிதரவில் எவ்வாறு சொற்களை சொல்வகை அடையாளப்படுத்தலாம் என்பதற்கான ஒரு எடுத்துக்காட்டைப் பார்ப்போம்.

வளாகத் <NNC> <B-NP>

தேர்வில் <NNC> <I-NP>

வேலைவாய்ப்பு <NN> <B-NP>

பெற்ற <VNAJ> <B-VNP>

மாணவர்களின் <NN> <B-NP>

பட்டியல் <NN> <I-NP>

வெளியிடும் <VNAJ> <B-VNP>

விழா <NN> <B-NP>

திங்கள்கிழமை <NNP> <B-NP>

நடைப்பெற்றது<VF> <B-VFP>

. <DOT> <O>

6.6.2.4 IIIT ஹைதராபாத்தில் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாட்டு அடையாளக் குழுமம்

இந்திய மொழிகளுக்கிடையிலான மொழிபெயர்ப்பிலும், ஆங்கில மொழியிலிருந்து இந்திய இயந்திர மொழிப்பெயர்ப்பிலும் சூடுப்பட்டுள்ள ஷைதராபாத்தில் உள்ள IIIT நிறுவனம் இந்தியாவில் இயற்கை மொழி ஆய்வுக்கு முன்னோடியாக உள்ளது. இந்நிறுவனம் உருவாக்கிய இந்திய மொழிகளுக்கான (குறிப்பாக இந்திக்கான) சொல்வகைப்பாட்டு அடையாளக் குழுமம் பரவலாகப் பயன்படுத்தப்படுகின்றது. இதன் பட்டடியல் கீழே தரப்பட்டுள்ளது.

NN	Common Noun
NNP	Proper Noun
*C	for all Compounds
QC	Cardinal No
QO	Ordinal No
PRP	Pronoun
VF	Verb finite main
VNF	Verb Non-finite adverb
VAUX	Verb Auxiliary

VNN Gerund/Verb non-finite nominal

VINF Verb Infinitive

JJ Adjectives

RB Adverbs

PSP Post position

QF Quantifiers

WQ Question words

CC Conjunction

NEG Negative

INT Interjection

CL Classifier

SVM Special

6.6.2.5 AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு

AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீட்டுப் பட்டியல்கள் கீழே தரப்பட்டுள்ளது. இப்பட்டியல் நிறுவன அறிவியலாளர் எல். சோபா என்பவரால் உருவாக்கப்பட்டது.

NOUNS

AU-KBC	CIIL	IIT-H
	NN - Common Noun	NN - Common Noun
	NNP - Proper Noun	NNP - Proper Noun
	NC - Noun Compound	*C - for all Compound
	CRD - Cardinal No	QC - Cardinal No
Ord - Ordinal No	ORD - Ordinal No	QO - Ordinal No
N.Sg - Noun Sg		

N.Pl - Noun Pl		
	PRP – Personal pronoun	PRP –Pronoun
	PRI – Idenfinite Pronoun	
	PRR – Reflexive Pronoun	
Pr.sg – Pronoun Sg		
Pr.Pl – Pronoun Pl		
Pr.Obl.Sg – Pronoun Oblique Sg		
V – Verb	VF – Verb Finite Main	VF - Verb Finite Main
	VNF – Ver non finite Adverbial adjectival	VNF – Ver non finite Adverbial adjectival
Vbn – verbal noun untensed, Vbn – Vbn Tensed Positive, Vbn Neg – Vbn Tensed Negative	VNN – Gerund/Verb non-finite nominal	VNN – Gerund/Verb non-finite nominal
Inf – Infinitive	VINF – Verb Infinitive	VINF – Verb Infinitive
VC – Verb causative	VCC – Verb Causative	
	VCD – Verb Double Cousative	
VT – Verb Transitive		
VI – Verb Intransitive		

ADJECTIVES

AU-KBC	CIIL	IIIT-H
Adj – Adjective		JJ – Adjective

	ADD – Adjective Declinable ADI – Adjective Indeclinable	

ADVERBS

AU-KBC	CIIL	IIIT-H
Adv – Adverb	ADV – Adverb ADL – Adverb of Location	RB – Adverb

OTHERS

AU-KBC	CIIL	IIIT-H
PSP – Postposition	PPO –Post position	PSP – Post position
Q – Quantifiers	QTF – Quantifiers	QF – Quantifiers
CNJ – Conjunction	CNJ - Conjunction	CC – Conjunction
	INT – Intensifer	INT – Intensifer
	NEG – Negative	NEG – Negative
	INJ - Interjunction	INJ – Interjunction
	PRT – Particle	RP – Particle

OTHERS

CIIL

IND – Indeclinable
QOT – Quotative
RDP – Reduplication
LWD – Loan Word
IDM – Idion
PRO – Proverb

இங்கு சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் பற்றிய முந்தய முயற்சிகள், சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பாக விளங்குகிறது. அவை CIIL சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் AUKBC நிறுவனத்தின் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாட்டு அடையாளக்குழுமம் AUKBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு குறிக்கப்பட்டுள்ளது. இவைகளை தேவை அடிப்படையில் பயன்படுத்த தேர்ந்தெடுத்துக் கொள்ளலாம்.

6.6.2.6. வாசு அரங்கநாதனின் டேக் தமிழ்

வாசு அரங்கநாதனால் உருவாக்கப்பட்ட டேக் தமிழ்ச் (Vasu Ranganathan's Tag Tamil) சொல்-ஒலியனியல் (Lexical-phonology) அணுகுமுறை அடிப்படையில் உருவாக்கப்பட்டது. டேக் தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வை உருபன்களின் வருகை முறையை வரிசை அட்டவணையைப் பயன்படுத்தி செய்கினறது. டேக் தமிழ் சொல் வகை அடையாளப்படுத்தலையும் உருவாக்கலையும் செய்கிறது.

6.6.2.7. கணேசனின் சொல்வகை அடையாளப்படுத்தி

கணேசன் தமிழுக்கு வேண்டி ஒரு சொல் வகை அடையாளப்படுத்தியை உருவாக்கியுள்ளார். அவருடைய அடையாளப்படுத்தி இந்திய மொழிகளின் மைய நிறுவன தமிழ்

விரிதரவில் நன்கு வேலை செய்கின்றது. பிற தரவுத்தொகுதிகளின் இதன் திறன் மதிப்பீடு செய்யப்பட வேண்டும். இவர் தமிழுக்கு வேண்டி விரிவான சொல்வகை அடையாளப்படுத்திகளின் குழுமத்தை வைத்திருக்கின்றார். இவர் இந்திய மொழிகளின் மைய நிறுவன தரவுத்தொகுதியின் ஒரு பகுதியை அகராதியாலும் உருபனியல் பகுப்பாய்வியாலும் சொல் வகைக்காக அடையாளப்படுத்தி உள்ளார். பின்னர் இவை அடையாளப்படுத்தப்பட்ட விரிதரவை மனிதசக்தியால் திருத்தி மீதியுள்ள பெருந்தரவுக்கு பயிற்சி தந்துள்ளார். இவருடைய சொல்வகை அடையாளப்படுத்தியில் சொல்வகை அடையாளப்படுத்தல் ஒவ்வொரு உருபன்களுக்கும் தரப்படுகின்றது.

6.6.2.8. RCILTS-இன் தமிழ் கதம்பம்

கதம்பம் ஆவணத்தில் தரப்பட்டுள்ள சொல்லுக்குச் சொல்வகை அடையாளத்தைத் தருகிறது. இது தமிழ்மொழியியல் அடிப்படையிலான ஊக விதிகளைப் பயன்படுத்துகின்றது. அகராதியையோ உருபனியல் பகுப்பாய்வியையோ பயன்படுத்துவதில்லை. இது பெரிய ஆவணங்களில் 80% திறனைத் தருகிறது. இது பன்னிரண்டு ஊக விதிகளைப் பயன்படுத்துகிறது. இது இட-எண்-பால், காலம் மற்றும் வேற்றுமை குறியீடுகளின் அடிப்படையில் சொல்வகை அடையாளங்களை அடையாளம் காண்கிறது. துனியாக நிற்கும் சொற்கள் சொல்வகை அடையாளப்படுத்தியில் சேகரிக்கப்பட்ட பட்டியலில் இருந்து பரிசோதிக்கப்படுகிறது. இது அறியப்படாத சொற்களுக்கு (unknown words) நிரப்பு விதியைப் பயன்படுத்துகிறது. இது bபைசயஅ மாதிரிகளைப் பயன்படுத்தி முன்வரும் சொல்லின் சொல்வகைப்பாட்டைக் கொண்டு அறியப்படாத சொல்லை அடையாளம் காண்கிறது.

6.6.2.9. பிட்ஸ் சொல்வகைப்பாடு அடையாளக் குழுமம்

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 குழுமமாகு. இவ்வடையாளக் குழுமம் 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

6.6.2.9.1. இ.இ.இ.மொ. ஒழுங்குமுறையில் பிட்ஸ் சொல்வகைப்பாட்டு அடையாளக் குடும்பம்

பிட்ஸ் சொல்வகைப்பாடு அடையாளக் குழுமம் இந்தியமொழியிலிருந்து இந்தியமொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையில் (இ.இ.இமொ ஒழுங்குமுறை) பயன்படுத்தப்படுகின்றது. தமிழுக்கும் இச்சொல்வகைப்பாட்டு அடையாளக்குழுமம் பின்பற்றப் படுகின்றது.

இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறையில் உட்படுத்தப்பட்டுள்ள எல்லா மொழிகளுக்கும் ஒரு பொதுவான சொல்வகைப்பாட்டு அடையாளக் குழுமம் பயன்படுத்தப்படுகின்றது. இச்சொல்வகைப்பாடுக் குழுமம் தான் தமிழுக்கும் பயன்படுகின்றது. சொல்வகைப்பட்டு அடையாளக் குழுமத்தில் பயன்படுத்தப்பட்டுள்ள அடையாளங்கள் கீளே விளக்கப்பட்டுள்ளன.

1. NN (Noun, பெயர்)

NN என்ற அடையாளம் இலக்கணத் தகவல்கள் அடிப்படையில் வேறுபடுத்தப்படாமல் பொதுவான பெயர்களுக்குப் பயன்படுத்தப்படுகின்றது.

எ.கா. நல்ல <JJ> கோவில் <NN>

2. NST (Noun denoting spial and temporal expressions, இட மற்றும் காலத்தைச் சார்ந்த பெயர்கள்)

முன்னால், *பின்னால்*, *அப்புறம்* போன்ற இட மற்றும் காலம் சார்ந்த சொற்கள் NST எனஅடையாளப் பயன்படுத்தப்படும்.

கோவிலின் <NN>முன்னால் <NST> நிற்கிற ஆள்

3. NNP (Pronoun, இயற்பெயர்)

எல்லா இயற்பெயர்களும் NNP என அடையாளப்படுத்தப்படும்.

4. PRP (Pronoun, மாற்றுப்பெயர்)

எல்லா மாற்றுப்பெயர்களும் PRP என அடையாளப்படுத்தப்படும்.

சென்னையிலிருந்து <PRP> மகாபலிபுரத்திற்கு <PRP> இருபது <QC> கிலோமீட்டர்கள் <NN>.

5. DEM (Demenostrative, சுட்டு)

எல்லாச் சுட்டுக்களும் DEM என அடையாளப்படுத்தப்படும்.

அந்தக் <DEM> கோவில் <NN>

6. VM (Verb main, முதன்மை வினை)

எல்லா முதன்மை வினையளும் VM என அடையாளப்படுத்தப்படும்.

அவர்கள் <PRP> சுற்றுலா <NN> சென்றனர்< VM. <SYM>

7. VAUX (Auxiliary Verb, துணைவினை)

எல்லாத் துணைவினைகளும் VAUX என அடையாளப்படுத்தப்படும்.

அவர்கள் <PRP> சென்னையிலிருந்து <NNP> மகாபலிபுரம் <NNP> செல்ல <VM> புறப்பட்டுக் <VM> கொண்டிருந்தனர் <VAUX>. <SYM>

8. JJ (Adjecive, பெயரடை)

பெரிய <JJ> கோவில் <NN>

9. RB (Adverb, வினையடை)

எல்லா முறைமை வினையடைகளும் RB என அடையாளப்படுத்தப்படும்.

பஸ் <NN> வேகமாகச் <RB> சென்று <MV>கொண்டிருந்தது <VAUX>. <SYM>

10. PSP (Postposition, பின்னுருபு)

எல்லா பின்னுருபுகளும் PSP என அடையாளப்படுத்தப்படும்.

அவர்களுடன் <PRP> கூடவே <PSP> வழிகாட்டியும் <NN> சென்றார் <VM>. <SYM>

11. CC (Conjucts, both co-ordinative and subordinative, இணைப்புகள், சமநிலை இணைப்பும் துணைநிலை இணப்பும்)

சமநிலை இணைப்புகளும் துணைநிலை இணைப்புகளும் CC என அடையாளப்படுத்தப்படும்.

கோயில்கள் <NN> மற்றும் <NN> சர்ச்சுகள் <NN> வழியில் <NN> இருந்தன <VM>.

12. WQ (Question Words, கேள்விச் சொற்கள்)

என்ன, எங்கே, எப்பொழுது போன்ற கேள்விச் சொற்கள் WQ என அடையாளப்படுத்தப்படும்.

13. QF (Quatifiers)

எல்லா அளவை அடைகளும் QF என அடையாளப்படுத்தப்படும்.

பல <QF> கோவில்களைக் <NN> கண்டு <VM>மகிழ்ந்தனர் <VM>.

14. QC (Cardinal, அடிப்படை எண்)

எல்லா அடிப்படை எண்களும் QC என அடையாளப்படுத்தப்படும்.

ஒன்று, இரண்டு, பத்து, நூறு போன்ற எல்லா அடிப்படை எண்களும் QC என அடையாளப்படுத்தப்படும்.

இரண்டு <QC> கிலோமீட்டர் <NN> தூரத்தில் <NN>

15. QO (Codinal, முறைமை எண்)

எல்லா முறைமை எண்களும் QO என அடையாளப்படுத்தப்படும்.

சென்னையில் <NNP> காண <VM> வேண்டிய <VAUX> இரண்டாவது <QO> இடம் <NN> மெரீனா <NNP> கடற்கரை <NN>ஆகும் <VAUX>. <SYM>

16. INFT (Inensifier, மிகையடை)

மிகையடை மொழியின் பெயரடைகளையுடனும் வினையடைகளுடனும் பயன்படுத்தப்படுகின்றது. எல்லா மிகையடைகளும் INFT என அடையாளப்படுத்தப்படும்.

மிகப் <INTF> பெரிய <JJ> கோபுரம் <NN>

பஸ் <NN> மிக <JJ> வேகமாகச் <RB> சென்றது <VM>. <SYM>

17. NEG (Negation, எதிர்மறை)

இல்லை போன்ற எதிர்மறைச் சொற்கள் NEG என அடையாளப்படுத்தப்படும்.

அங்கு <NST> குளிப்பதற்கு <VM> இடம்<NN> இல்லை <NEG>. <SYM>

18. SYM (Speical Symobl, சிறப்புக் குறியீடு)

பிற அடையாயாளங்களால் பாகுபடுத்தப்பட இயலாத சொற்கள் SYM என் அடையாளப்படுத்தப்படும். $ போன்ற சிறப்புக் குறியீடுகள் SYM என அடையாளப்படுத்தப்படும்.

19. RDP (Reduplication, இரட்டிப்பு)

இந்திய மொழிகளில் சில சொற்கள் இரண்டு தடவை எழுதப்படுகின்றன. அம்மாதிரியான சொற்கள் RDP என அடையாளப்படுத்தப்படும்.

யாத்திரிகள் <NN> மலையில் <NN> மெல்ல மெல்ல <RDP> அடிவைத்து <MV> ஏறினார்கள் <VM>. <SYM>

20. ECH (Echo Words, எதிரொலிச் சொல்)

எதிரொலிச் சொற்கள் இந்திய மொழிகளில் பொதுவாகக் காணப்படும்.

அவர்கள் <PRN> தடதட <ECH> என்று படிகளில் <NN> ஏறினார்கள் <VM>.

21. UNK (Unkown Words, தெரியாத சொற்கள்)

தெரியாத சொற்கள் UNK என அடையாளப்படுத்தப்படும்.

6.6.2.9.2. சொல்வகைப்பாடு அடையாளப்படுத்தும் பொறியின் செயல்பாட்டுத் தனிக் குறிப்பீடுகள் (Functional Specifications of POS tagging Engine)

வாக்கியங்களில் வரும் ஒவ்வொரு சொல்லுக்கும் சொல்வகை அடையாளப்படுத்தும் செயல்பாடு சொல்வகை அடையாளப்படுத்தல் (Parts of Speech (POS) tagging) எனப்படும். வாக்கியங்களின் வரும் உறுப்புகளின் பங்களிப்பை ஆய்வதற்குப் பெயர், வினை, பெயரடை, வினையடை போன்ற சொல்வகைப்பாடுகளைக் கண்டுபிடிப்பது உதவும். இதற்கு விதி அடிப்படையிலானது (rule-based), புள்ளியியல் அடிப்படையிலானது (statistics based), மாற்றம் அடிப்படையிலானது (transformation based) எனப் பல அணுகு முறைகள் இருக்கின்றன. இ-இ இயந்திர மொழிபெயர்ப்பில் புள்ளியியல் அணுகு முறை பயன்படுத்தப்படுகின்றது.

6.6.2.9.2.1 இடுவரல் விடுவரல் தனிக்குறிப்பீடுகள்

Input TKN_

Output CAT_

CAT_ என்ற தனிதன்மையின் எல்லாச் சாத்தியங்களான மதிப்புகளும் இங்கு பட்டியலிடப்படவேண்டும். NN, PRP, VM மற்றும் பிற.

இடுவரல் தனிக்குறிப்பீடுகள் TKN_ என்ற தனிதன்மை இடுவரல் சொல்வகைப்பாடு அடையாளப்படுத்திக்குத் தரப்படும் SSF-இல் வரையறை விளக்கம் செய்யப்படவேண்டும் என வேண்டும்.

விடுவரல் தனிக்குறிப்பீடுகள் CAT_ என்ற தனித்தன்மை சொல்வகைப்பாடு அடையாளப்படுத்தியால் வரையறை விளக்கம் செய்யப்படவேண்டும் என வேண்டும். எனவே விடுவரல் SSF சில செல்லுபடியாகும் மதிப்புகளுக்கு CAT தனிப்பண்பைக் கொண்டிருக்கவேண்டும்.

CAT_ என்பது PRP, NN போன்ற மதிப்புகளைக் கொண்டிருக்கும்.

எல்லா CAT_ மதிப்புகளையும் இங்கு பட்டியலிடவும்; பின்னர் ஒவ்வொன்றிற்கும் ஒரு சுருக்கமான வருணனையைத் தரவும்.

PRP - Pronoun

NN – Noun

எடுத்துக்காட்டு:

Input

ADDR_	TKN_	OTHR
1	Afwa	<fs af =’ afwa pronn…’>
2	pEyanE	<fs af =’ pEyanE, noun…’>
3	Pola	<fs af =’ pola, psp…’
4	irukkirYAn	<fs af =’ irukkiRaan, verb…’

Output

ADDR_	TKN_	CAT_	OTHR
1	Afwa	PRP	<fs af =’ afwa, pron…’
2	pEyanE	NN	<fs af =’ pEyanE, noun…’
3	Pola	PSP	<fs af =’ pola, psp…’>
4	irukkirYAn	Verb	<fs af =’ irukkirYAn, verb…’

(குறிப்பு: நீங்கள் ஒரு பண்புக்கூறை (feature) பயன்படுத்துவதானால் இடுவரல் மாதிரியிலும் விடுவரல் மாதிரியிலும் காட்டவும்.)

6.6.2.9.2.2 சொல்வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம்

சொல்வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம் ஒரு தொகுதியிலிருந்து மற்றொரு தொகுதிக்கு வழியமைப்புக் கட்டுப்பாட்டு ஒழுக்கின் (program control flow) முழுக் காட்சியையும் காட்டும். இது வழியமைப்புக்குள் பல்வேறுபட்ட தீர்மான இடங்களைக் காட்டும்.

6.6.2.9.2.3 செயல்பாட்டு வருணனைகள்

- SSF வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று (Covert SSF to TnT Format)
- பயிற்சிக் கோப்பை SSF வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று. மேலும் சோதனை கோப்பையும் வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று.
- TnT வடிவமைப்பிலிருந்து SSF வடிவமைப்புக்கு மாற்று (Convert TnT Format to SSF)
- TnT வடிவமைப்பில் உள்ள சொல்வகைப்பாடு அடையாளப்படுத்தியால் உருவாக்கப்பட்ட இடுவரலை SSF வடிவமைப்புக்கு மாற்று.
- எழுத்துக் கிளையை உருவாக்கு (Build Letter Tree)
- முன்னொட்டு கிளைத் தரவு அமைப்பை உருவாக்கப் பயிற்சிக் கோப்பிலிருந்து சொற்களையும் அடையாளங்களையும் எடுக்கவும். கிளையமைப்பில் சொல்லையும் அடையாள நிகழ்வெண்ணையும் சேகரிக்கவும். சொல்லையும் அதன் நிகழ்வெண்ணையும் பங்கெடுப்பாளர்களாகக் எடுத்துகொண்டு எழுத்துக் கிளையை உருவாக்கவும்.
- புடைபெயர்வு எண்ணிக்கை சட்டத்தைக் உருவாக்கவும் மற்றும் வெளிவரு எண்ணிக்கை சட்டத்தையும் உருவாக்கவும் (Build Transtition Count Matrix and Build Emission count matrix)
- அடையாள வரிசை மற்றும் அதன் நிகழ்வெண்ணின் கலவையை உருவாக்கவும்.

- மிகச் சிறந்த அடையாள வரிசையைக் கண்டுபிடித்தல் (Finding Best Tag Sequence)
- ஒரு வாக்கியத்தை எடுக்கவும் மிகச் சிறந்த அடையாள வரிசையைக் கண்டுபிடிக்க Veterbi algorithm-ஐப் பயன்படுத்தவும்.
- அடையாள வரிசையைக்கு மெருகிடுதல் (Smoothing for Tag sequence)
- ஒரு அடையாள வரிசை இல்லாவிட்டால் சொல்வகைப்பாட்டின் பங்கெடுப்பாளர் அடிப்படையில் மெருகிடும் தொழில்நுட்பத்தைப் பயன்படுத்தவும்.
- சொற்களுக்கு மெருகிடுதல் (Soothing for Words)
- சட்டகங்களில் சொல் கண்டுபிடிக்கப்படாவிட்டால் பின்னொட்டு மெருகிடும் தொழில்நுட்பத்தைப் பயன்படுத்தவும்
- N-gramms-க்கு மெருகிடுதல் (Soothing for N-grams)

6.6.2.10. சர்வேஸ்வரன் மற்றும் மகேசன் என்போரின் விதி அடிப்படையிலான படிநிலை அடையாளக் குழு

சர்வேசரன் மற்றும் மகேசன் என்போர் விதி அடிப்படையிலான படிநிலை அடையாளக் குழு (Hierarchical Tag-set for Rule-based Processing of Tamil Language) முன் மொழிகின்றனர் (Sarveswaran and Mahesan 2014. இவர்கள் பெயர்ச்சொல், வினைச்சொல், முற்றுச்சொல், பகுக்கவியலும் சொல், கூட்டுச்சொல், கடன்சொல், குறியீடு, நிறுத்தற்குறி, பிற, தெரியாதன என பத்து சொல்வகைப்பாட்டு அடையாளக் குழுவை முன்மொழிகின்றனர்.

6.7. முடிவுரை

தரவுத்தொகுதி மொழியியலில் சொல்வகைப்பாடு அடையளப்படுத்தல் என்பது ஒரு உரையில் (அல்லது தரவுத்தொகுதியில்) ஒரு சொல்லை அதன் வரையறைவிளக்கம், அதன் வடிவம் மற்றும் அது வரும் சூழல் அடிப்படையில் அதை இலக்கண வகைப்பாட்டிற்கு அடையாளபடுத்தும் செயல்பாடாகும். ஒருசொல் அது ஏற்கும் இலக்கணக்கூறுகள் அடிப்படையிலும் அது ஒருவாக்கியத்தில் வரும் சூழ அடிப்படையிலும் இலக்கண வகைப்பாடு செய்யப்படுகின்றது. இந்த வகைப்பாடு ஒரு வாக்கியதில் அதன் செயல்பாட்டை (function) குறித்து நிற்கும். ஒரு சொல்லின் இலக்கண வகைப்பாடு அச்சொல்லின் உருபனியல் பண்பையும் தொடரியல் பண்பையும் உணர்த்தி நிற்கும். ஒரு சொல்லை அது ஒரு வாக்கியத்தில் வரும் அர்த்தத்தைப் புரிந்துகொள்வதற்கு அதன் வகைப்பாடு மிகவும் முக்கியமாகும். எனவே இயந்திர

மொழிபெயர்ப்புக்கு உள்ளாக்கப்படும் உரை அல்லது தரவுத்தொகுதி சொல்வகைப்பாட்டிற்காக அடையாளப் படுத்தப்படவேண்டும். சொல்வகை அடையாளக் குழுமத்தின் உறுப்பினர்களின் எண்ணிக்கை சொல்வகை அடையாளப் படுத்தப்படுவதன் குறிக்கோள் அடிப்படையில் மாறும். ஒரு கூட்டம் சொல்வகைப்பாடு உறுப்பினர்கள் ஒரு சொல்வகைப்பாட்டுக் குழுமத்தை உருவாக்கும். முன்னர் கூறியது போல் இதன் எண்ணிக்கை வேறுபடும். சொல்வகை அடையாளக் குழுமத்தைத் தட்டையான அடையாளக்குழுமம் என்றும் படிநிலையான அடையாளக் குழுமம் என்றும் பகுப்பர்.

இயல் 7

தொடரியல் பகுப்பாய்வு

7.1. அறிமுகம்

சொல்வகை அடையாளப்படுத்தலுக்குப் பின்னர் செய்ய வேண்டிய செயல்பாடு சொற்களை தொடர்களாகக் குழுதல் ஆகும். அதாவது வாக்கியங்களைத் தொடர்களாகப் பகுத்தலாகும். இதைத் தொடரியல் பகுப்பாய்வு (syntactic parsing) எனலாம். இலக்கணப் பகுப்பாய்வு என்பது ஒரு இலக்கண அடிப்படையில் பனுவல்களின் தானியக்க ஆய்வுகளுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு பனுவலுக்கு தொடரியல் அமைப்பைத் தரும் செற்பாங்கைக் குறிப்பிடுகின்றது. இது பெரும்பாலும் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுப்பிடிக்கப்பட்ட பின் நடைமுறைப்படுத்தப்படும். வேறுபட்ட இலக்கணங்கள் (சார்பு இலக்கணம், சூழல் இலக்கணம் போன்றவை) அடிப்படையில் பகுப்பாய்வு செய்தல் ஒன்றுடன் ஒன்று உயர்ந்த நிலைத் தொடரியல் உறவுகளில் கொண்டு வருகிறது. வாக்கிய நிலைப் பகுப்பாய்வு சொல்நிலைப் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்த தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உள்ளடக்குகின்றது. ஒரு பகுத்து அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பகுத்துக் குறித்தலில் கிளைப்படங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree bank) என்றழைக்கப்படுகின்றது. கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் அடையாளப்படுத்தப்பட்ட அடைப்புக்குறிகளின் குழுமத்தைப் பயன்படுத்தி உருப்படுத்தம் செய்யப்படும்.

எ.கா.

‘Pearl sat on a chain’ என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்.

[S [NP Pearl – NP/NP] [VP – Sat VVI] [PP on II]

[NP a – AJI] [Chair – NNI]

[NP] [PP] [VP] S]

இதில் உருபனியல், தொடரியல் தகவல் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்பட்டிருக்கும் உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் திறக்கும் மற்றும் அடைக்கும் செவ்வக அடைப்புக் குறிகளால் அடையாளப்படுத்தப்படும்.

எல்லாப் பகுத்துக் குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருமாறு:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை.

2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வகைகள்

இந்த வேறுபாடுகள் இருப்பினும் பெரும்பாலான பதுத்துக் குறித்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையால் அமைந்தது.

இந்த அமைப்பொழுங்குக்குள் முழு பகுத்துக் குறித்தல் [full parsing] திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாகக் கொண்டது. குறுகிய பகுத்துக் குறித்தல் திட்டம் (Skeleton parsing scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தலைப் பயன்படுத்துகின்றது. மேலும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகின்றது. பகுத்துக் குறித்தல் பெரும்பாலும் மனித ஆய்வில் பின் திருத்தம் செய்யப்படுகின்றது. ஏனென்றால் தானியங்கு பகுத்துக் குறித்தல் சொல்வகை அடையாளப்படுத்தலை விட வெற்றி குறைந்தது. முழு மனித இயக்கப் பகுத்துக் குறித்தலின் குறைப்பாடு பகுத்துக் குறித்தல், தரவுத்தொகுதி திரித்தல் இவற்றில் ஈடுபடும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுக்கட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள் கோள் சாத்தியமாகும்போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் வேறுபட்ட நிலைகளில் (சொல்நிலை, வாக்கியநிலை, தொடர்நிலை, செயல்பாடு–பங்கெடுப்பாளர் நிலை) அடையாளப்படுத்தும் நிலை தரும் மொழி மூலவளமாக அமைகின்றது. கிளை வங்கிகள் இயற்கைமொழி ஆய்வு தரவு உந்தல் அடிப்படைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுத்தல், பொதுவாக மொழியியல் ஆய்வுகள் இவற்றின் உருவாக்கத்திற்கு முக்கியமானதாகும்.

அடிப்படையில் தொடர் பகுப்பான் (chunker) வாக்கியத்தை மேலுறல் செய்யாத முக்கியமான தொடர்களாகப் பிரித்து அதற்குப் புலக்குறிப்பு (label) தருகிறது. தொடர்பகுப்பான்

அவற்றின் துல்லியமான வெளிப்யீடு அடிப்படையிலும் எவ்வாறு தொடர் பகுதிகள் விளக்கப்படுகிறது என்பதன் அடிப்படையிலும் வேறுபடுகின்றன. பல தொடர் பகுப்பான்கள் எளிய தொடாகளைப் பகுத்தலைவிடக் கூடுதலாகச் செய்கிறது. பிற பெயர்த்தொடர்களை மட்டும் கண்டுபிடிக்கிறது. தொடர்பகுப்பான்கள் சொல்வகை அடையாளப்படுத்தலுக்கும், முழு பகுத்துக் குறித்தலுக்கும் இடையில் வருகிறது. தனிப்பட்ட தொடர் பகுதிகளின் அமைப்பை விளக்குவது எளிது. ஆனால் அவற்றிற்கிடையிலான உறவுகளை விளக்குவது கடினம். இது தனிப்பட்ட சொல் வகுப்புகளைச் சார்ந்திருக்கிறது. இவ்வாறு தொடர்களைப் பகுத்தில் தற்பொழுது இருக்கின்ற முறைக்கும் முன்மாதிரி வெளிப்யீட்டு ஆய்வுக்கும் இடையில் உள்ள சலுகையான முறையாகும். தொடர் பகுப்பான்கள் வாக்கியங்களைப் பிரித்து அடையாளப்படுத்துகிறது. பெரும்பாலான தொடர் பகுப்பான்கள் சொல்வகை அடையாளப்படுத்தலில் உள்ள தகவலைப் பயன்படுத்துகிறது. பிற உண்மையான சொற்களைப் பார்க்கிறது.

சொல் வகைப்பாட்டு அடையாளப்படுத்தி சொல் மற்றும் சூழலுக்கான தகவலைப் பயன்படுத்தி இம்மயக்கங்களை நீக்குகின்றது. சொல் வகைப்பாட்டுத் தகவல் பகுத்துக்குறித்த அமைப்பில் உறுப்பின் நிரல் அல்லது இடத்தைக் கண்டுகொள்ள தேவைப்படுகிறது. பண்புக்கூறு தகவல்கள் நிரப்பியையும் உடன்பாட்டையும் பரிசோதிக்கப் பயன்படுகிறது.

தமிழில் ஒத்தறி அடிப்படையில் சொல்நிரல் சுதந்திரமானது. தொடரமைப்பு விதிகளின் எல்லா விதமான வரிசை மாற்றங்களுக்கும் வேண்டி எழுதப்பட வேண்டும் என்று அவசியப்படும் இவ்வகங்களின் காரணமாக இப்பகுத்துக் குறிப்பான்கள் தேவையில்லாமல் விதிகளின் எண்ணிக்கையைப் பெருக்கிக் கொள்ளும். பகுத்துக்குறிப்பான் மீது கூடதலான சுமை புகுத்தப்படும். தற்போதைய ஆய்வில் பகுத்துக்குறித்தலுக்கு முன் வாக்கியத்தை ஒரு குறிப்பிட்ட நிலை பேறாக்கம் செய்யும் நடத்தை பின் பற்றப்படுகிறது.

வினைதான் வாக்கியத்தின் முக்கிய உறுப்பாகும். வினைகள் அவற்றின் பொருண்மை இயல்பு அடிப்படையில் வேறுபட்ட பங்களிப்பாளர் அமைப்பை ஏற்கும். இப்பங்கெடுப்பாளர் அமைப்புகள் பெயர்த்தொடர்களின் வேற்றுமை உருபுகளை விளக்கும். அவை வினைக்கு நேரடியான மற்றும் நேரடியல்லாத செயப்படுபொருளாக அமையும். பகுத்துக்குறிப்பான் வாக்கியத்தில் வினைக்குப் பொருத்தமான பங்கெடுப்பாளர் அமைப்பை பரிசோதிப்பதைக்

கவனிக்க வேண்டும். மேலும், தமிழில் எழுவாயின் பால்-எண்-இட பண்புக்கூறுகள் பயனிலையுடன் இயைபுடையதாக இருக்க வேண்டும். ஒரு பகுத்துக்குறிப்பான் வாக்கியத்தைப் பகுத்துக்குறிக்கும்போது வாக்கியத்தைச் சரியாக மதிப்பிட இவ் இயைபு பண்புக்கூறைப் பரிசோதிக்க வேண்டும்.

7.2 இலக்கணப் பகுப்பாய்வு (parsing)

இலக்கணப் பகுப்பாய்வு இயந்திரமொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுத்தல் போன்ற பயன்பாடுகளில் முக்கியமான செயல்பாடாகும். சில பயன்பாடுகள் வாக்கியங்களின் முழுமையற்ற பகுத்துக்குறித்தலை (shallow parsing) வேண்டுகிறது. ஆங்கிலம் மற்றும் பிற ஐரொப்பிய மொழி வாக்கியங்களின் பகுத்துக்குறித்தலுக்குப் பல முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இந்திய மொழிகளில் இந்தி மற்றும் வங்காள மொழிகளுக்கு பகுத்துக்குறிப்பான்கள் இருந்தாலும் தமிழைப் பொறுத்தவரையில் கூறத்தக்க பகுத்துக்குறிப்பான் இல்லை. தமிழில் பல கோட்பாட்டு மொழியியல் அடிப்படையிலான தொடரியல் ஆய்வுகள் செய்யப்பட்டுள்ளன. இருப்பினும், தமிழ் வாக்கியங்களின் பகுத்துக்குறித்தலுக்கு வேண்டிய கணினியியல் சார் சட்டகத்தை உருவாக்க அச்சிட்டு வெளிப்யிடப்பட்ட புத்தகங்கள் இல்லை. தற்போதைய ஆய்வு தமிழ் வாக்கியங்களின் பகுத்துக் குறித்தலுக்கு கணினி அடிப்படையிலான இலக்கணச் சட்டகம் உருவாக்குவதில் கவனம் செலுத்தப்படுகின்றது.

7.3 ஆழமிலாப் இலக்கணப் பகுப்பாய்வு (shallow parsing)

நாம் தொடரியல் என்ற கலைச்சொல்லை மரபுப் பகுப்பாய்விலிருந்து கிடைக்கும் வெளிப்யீட்டை விடக் குறைவான நிறைவுடைய ஆய்வைக் குறிக்கும் பொதுவான சொல்லாகப் பயன்படுத்துகிறோம். ஆய்விலிருந்து கிடைக்கும் வெளிப்யீடு தொடரமைப்புக் கிளைப்படம் அல்ல. ஒரு பகுப்பாய்வி பெயர்த்தொடர்கள் போன்ற பகுக்கப்பட்ட உறுப்புகளை வாக்கியத்தில் அவற்றின் அக அமைப்பையும், செயல்பாட்டையும் குறிப்பிடாமல் அடையாளம் காணும். மற்றொரு வகையான ஆய்வு முதன்மை வினை மற்றும் அவற்றின் நேரடியான பங்கெடுப்பாளர் போன்ற சொற்களின் சில செயல்பாட்டுப் பங்களிப்பை அடையாளம் காணும். பகுப்பாய்வு ஒழுங்குமுறை உருபயனியல் ஆய்வின் மீதும் பொருண்மை மயக்க நீக்கலின் மீதும் இயல்பாக

வேலை செய்யும். இதன் அடிப்படை நோக்கம் தலைச்சொல் (Head word) உருபனியல் தகவல் மற்றும் சொல்வரிசை அமைப்பு இவற்றில் இருந்து சாத்தியமான தொடரியல் அமைப்பைப் (Syntantic structure) பெறுவதாகும். முன்மாதிரியாக பகுப்பாய்வுத் தொடர்களையும் தலை-அடை (head-medifier) உறவுகளையும் கண்டுபிடிப்பதை நோக்கமாகக் கொள்கிறது. பெரிய உரை தரவுத்தொகுதிகளின் (corpus) பயன்பாட்டில் ஆழமில்லாப் பகுப்பாய்வி பயன்படுத்தப்படுகிறது. பெரும்பாலும் பகுப்பாய்விற்கு எல்லாச் சிக்கல்களுக்கும் தீர்வு காண்பதற்கு போதிய திறன் இல்லாவிட்டால் பகுதியான ஆய்வுகள் அனுமதிக்கப்படும். சர்ச் (Church.1988) பொருத்தமான அடைப்புக் குறிகளைச் செருகி எளிய பெயர்த் தொடர்களை அடையாளம் காணும் புள்ளியல் (Stochastic) வழியமைப்பை வடிவமைத்துள்ளார். ஒப்னெய் (Obney, 1991) உளமொழியியல் சான்றுகள் அடிப்படையிலும், நடைமுறை பயன்பாடுகள் அடிப்படையிலும் ஆழமில்லாப் பகுப்பாய்வின் முக்கியத்துவத்தை முதலில் விவாதித்தவர் என்று மதிக்கத்தக்கவர் ஆவார். அவர் தமது அணுகுமுறையில் ஆழமில்லாப் பகுப்பாய்வைப் பெறுவதற்காக முற்றுநிலை மாறிகளைப் பயன்படுத்தினார். ஆழமில்லாப் பகுப்பாய்வி சொல்வகைப்பாடு அடையாளப்படுத்தல், தொடர் உறுப்பு வரையறுத்தல், தொடர் உறவுகளைக் கண்டுபிடித்தல் என்ற முன்மாதிரி தொகுப்புகளை உள்ளடக்கியது.

7.3.1 சொல்வகை அடையாளப்படுத்தல் (tagging)

ஒரு சொல்லும் அதன் சூழலும் தரப்படுகையில் சொல்வகை அடையாளப்படுத்தி அச்சொல்லின் சரியான உருபனியல்-தொடரியல் வகுப்பை (பெயர், வினை கொண்ட பகுப்பை) உறுதிசெய்யும் சொல்வகை அடையாளப்படுத்தி சொல்வகை அடையாளப்படுத்தல் (Parts of Speech tagging) இயற்கைமொழி ஆய்வில் மிக அறியப்பட்ட சிக்கலாகும். இதில் இயந்திர கற்றல் அணுகுமுறைகள் (machine language approaches) பொதுவாகப் பயன்படுத்தப்படுகிறது.

7.3.2 தொடர்க்கூறு பகுப்பாய்வு (chunking)

சொற்களும் அவற்றின் உருபனியல்-தொடரியல் வகுப்புகளும் தரப்படுகையில் எந்தச் சொற்கள் தொடர்கூறுகளாகக் (பெயர்த்தொடர், வினைத்தொடர், நிரப்பியத்தொடர்

போன்றவைகளாக) குழுமப்பட வேண்டும் என்று தொடர்கூறு பகுத்தல் (Chunking) செயல்பாட்டால் தீர்மானிக்கப்படுகிறது.

தொடர்கூறின் வரையறை விளக்கத்துடன் பின்வரும் கேள்வி தொடர்புள்ளது. தொடர் உறுப்பு எதைக் கொண்டிருக்கும்? ஒரு மாதிரி தொடர் உறுப்பு ஒரு தனி நிலைப் பொருளடக்கச் சொல்லைச் சுற்றி ஒரு கூட்டம் செயல் பாட்டுச் சொற்களைக் கொண்டிருக்கும் (Abney, 1991). தொடர் உறுப்புகள் இயல்பாக ஒன்றுக்கொன்று இயைபுள்ள சொற்களின் குழுமமாக எடுத்துக்கொள்ளப்படுகின்றது.

அடுத்த கேள்வி 'நம்முடைய தேவைக்காக இந்த இயைபுள்ள சொற்களின் குழுவின் எல்லைகளை எவ்வாறு வரையறை விளக்கம் செய்யவியலும்?' என்பதாகும். எடுத்துக்காட்டாக, பின்வரும் இணைகளில் எது சரியான தொடர் உறுப்பு? என்ற ஐயம்.

((மேசையின் மேல்)) அல்லது

((மேசையின்)) ((மேல்))

7.3.3 தொடர்கூறுகளின் உறவைக் கண்டுபிடித்தல்

ஒரு வாக்கியத்தில் உள்ள தொடர்கூறுகள் தரப்படுகையில் முக்கிய வினையுடன் அவற்றிற்குள்ள உறவுகள் (எழுவாய், பயனிலை, இடம் போன்றவை) தீர்மானிக்கப்பட வேண்டும் என்ற நிலையில் உறவைக் கண்டுபிடித்தல் (Relation finding) முக்கியமான செயல்பாடாகும்.

ஆழமில்லாப் பகுப்பாய்விகள் இயற்கை மொழி முழுவதையும் கையாள வேண்டி இருப்பதால் அவை பெரிதாகவும் பெரும்பாலும் ஆயிரக்கணக்கான விதிகளையும் கொண்டிருக்கின்றன. இவ்விதிகளில் பல விதிவிலக்கானவை என்பதன் அடிப்படையில் இவை வலிமை குறைந்தவையாக இருக்கின்றன. ஆழமில்லாப் பகுப்பாய்விகள் இயந்திர கற்றல் அடிப்படையிலான தொழில்நுட்பங்களைப் (techniques) பயன்படுத்தித் தானியக்கமாக உருவாக்கப் படுகின்றன.

7.4 பெயர்த்தொடர் பகுப்பாய்வு

பெயர்த்தொடர் உறுப்பு வரையறுத்தல் (பகுத்தல்) வாக்கியங்களிலிருந்து பெயர்த் தொடர்களைப் பிரித்தெடுப்பதைக் குறிப்பிடுகின்றது. பகுத்துக்குறித்தலைவிட பெயர்த்தொடர் பகுத்தல் எளிதாக இருந்தாலும் ஒரு துல்லியமான மற்றும் திறமைவாய்ந்த பெயர்த்தொடர்

பகுத்தல் பல பயன்பாடுகளில் பயன்படுத்தப்படுகின்றது என்ற உண்மையால் மிக முக்கியத்துவம் வாய்ந்ததாகக் கருதப்படுகின்றது.

பெயர்த்தொடர்களைப் பனுவல்களைப் பகுத்துக் குறிப்பதற்கு முன் முன்பகுப்பாய்வுக் கருவியாகப் (pre-processing tool) பயன்படுத்த இயலும். இயற்கை மொழியின் கூடுதலான பொருண்மை மயக்கம் காரணமாக பனுவலின் சரியான பகுத்துக்குறித்தல் மிகக் கலவைத் தன்மையானதாக (சிக்கலானதாக) இருக்கின்றது. இந்நேர்வுகளில் தொடர்களைப் பகுத்தலை இம்மயக்கங்களை ஓரளவு தீர்ப்பதற்கான முன்பகுப்பாய்வுக கருவியாகப் பயன்படுத்த இயலும். இப்பயன்பாட்டில் தொடர் பகுத்தலைச் சொற்களுக்குப் பதிலாகத் தொடர் பகுதிகளைச் சார்ந்திருக்கும் ஆவணங்களிலிருந்து தரவை மீட்பதற்குப் பயன்படுத்த இயலும். குறிப்பாகப் பெயர்களும் பெயர்த்தொடர்களும் தகவல் மீட்புக்கும் பிரித்தெடுத்தலுக்கும் மிகவும் பயனுள்ளதாகும். இயந்திர மொழிபெயர்ப்பு மீதான சமீபகாலச் செயல்பாடுகள் பயனுள்ள மாற்ற அமைப்பொழுங்குகளை (transfer patterns) ஆக்குவதற்கு இரு மொழிகளிலுள்ள பனுவல்களை (இணை தரவுத்தொகுதிகளை) (parallel corpora) பயன்படுத்துகின்றது. பெயர்த்தொடர்கள் இணை தரவுத்தொகுதிகளில் உள்ள பனுவல்களைப் பொருத்தும் பயன்பாடுகளில் பயன்படுகின்றது. இணை தரவுத்தொகுதிகளில் உள்ள வாக்கியங்களைத் தொடர்ப்பகுப்புத் தகவல்களைப் பயன்படுத்தியும் மூல மொழியிலுள்ள தொடர் பகுதிகளை இலக்கு மொழியிலுள்ள தொடர் பகுதிகளுடன் உறவுபடுத்தியும் பொருத்த வரைவு செய்யலாம். மேலும் பகுக்கப்பட்ட பெயர்த் தொடர்களைத் தரவின் தரமான பகுத்துக்குறித்தல் தேவைப்படாத பிற பயன்பாடுகளிலும் பயன்படுத்த இயலும்.

7.5. தமிழ்த் தரவுத் தொகுதியைத் தொடர்கூறுக்குப் பகுத்தல்

7.5.1 தமிழ்த் தொடரியல் அமைப்பு

தமிழைப் பொறுத்தவரையில் அது ஒரு ஒட்டுக்கள் நிறைந்த மொழியாகும். இவ்வொட்டுக்கள் வேர்ச்சொற்களில் சேர்க்கப்பட்டு வேறுபட்ட சொல் வடிவுகளை உருவாக்கும். ஏனவே தமழுக்கு ஒரு உருபனியல் ஆய்வியை உருவாக்குவது ஒரு சிக்கலான செயல்பாடாகும். பல வாக்கியங்களுக்கு உருபனியல் பகுப்பாய்வியைப் பயன்படுத்தி சொற்களின் அல்லது ஒரு சொல்லின் சொல் வகைப்பாட்டையும் (Parts of speech (POS)) பண்புக்கூறு தகவலையும்

அடையாளம் காண இயலும். இருப்பினும், சில சொற்களுக்கு மயக்கமுள்ள சொல்வகைப்பாடும் இலக்கண பண்புக்கூறுகளும் இருக்கும்.

தமிழ்த் தொடரியல் பகுத்துக் குறிப்பானுக்கு அடிப்படையாக அமையும் தமிழ்த் தொடரியல் குறித்து இங்கு விளக்கப்படும். தமிழ் வாக்கியங்களின் வகைகள் பற்றியும், அவற்றின் அமைப்பு குறித்தும் விளக்கப்படும்.

தமிழ்மொழி வேறுபட்ட சொல் வரிசை அமைப்பைக் கொண்ட தொடர்களைக் கொண்டு வாக்கியங்களை அமைக்கத்தக்க வகையில் அமைந்த மொழியாகும். சொற்களுக்கு இடையிலான தொடர்பினைப் பயன்படுத்தி வாக்கியங்களை இலக்கணரீதியாக அமைக்கிறோம். இலக்கணங்கள் வேறுபட்ட வார்த்தைகளுக்கு இடையே தொடர்பினை உருவாக்கிப் பொருள்தரும் வாக்கியங்களை உருவாக்க உதவுகிறது. இவ்வாறு சொற்களைக் கொண்டு தொடர்களையும் தொடர்களைக் கொண்டு வாக்கியங்களையும் உருவாக்கும் விதிகளைக் கூறுவது தொடரியலாகும்.

ஒன்றுக்கு மேற்பட்ட சொற்களஅத்தொடர்கள் இணைந்து உருவாகும் வாக்கியங்கள், அவ்வாக்கியங்கள் இணைந்து உருவாகும் கலப்பு, கூட்டு வாக்கியங்கள் ஆகியவற்றின் கட்டமைப்பையும் அவை தரும் பொருண்மையையும் ஆராய்வது தொடரியல் எனப்படும். பொதுவாகத் தொடர்களை ஆராய்வது தொடரியல். அது தொடராக இருக்கலாம் அல்லது வாக்கியங்களாக இருக்கலாம். வாக்கியம் என்பதும் ஒருவகைத் தொடரே.

7.5.1.1 தொடர்கள்

தமிழ் வாக்கியங்களைப் தலைமையாய் வரும் சொல்வகைப்பாடுகளின் அடிப்படையில் பெயர்தொடர், வினைதொடர், பெயரடைத் தொடர், வினையடைத் தொடர், பின்னுருபுத்தொடர் என வகைப்படுத்தலாம்.

7.5.1.1.1 பெயர்த்தொடர்

பெயர்த்தொடர் என்பது ஒரு பெயர்ச்சொல்லைத் தலைமைப் பெயராகக் கொண்ட தொடர் ஆகும். தலைமைச்சொல் பெயராகவோ ஆக்கப்பெயராகவோ (எகா.மயக்கம்) தொகையாகவோ (எகா. பணப்பெட்டி) வினைப்பெயராகவோ, மாற்றுப்பெயராகவோ (எகா.அவன், நான்,நீ போன்று) தொழில் பெயராகவோ (வந்தது, வருவது, வருதல், வந்தமை, வந்தவன் போன்று)

இருக்கலாம். பெயர்த்தொடரில் பெயருக்கு முன் அடைகொளி அடை, எண்ணடை, பெயரடை, அளவையடை போன்ற அடைகளும் பெயரெச்சத் தொடரும் வரலாம்.

எ.கா.

பையன், அந்தப் பையன், பெரிய பையன், இரண்டாவது பையன், இரண்டு பையன்கள், என்னுடைய பையன், நேற்று வந்த பையன்.

7.5.1.1.2 வினைத்தொடர்

ஒரு வினைச்சொல்லைத் தலைமைச் சொல்லாகக் கொண்டு உருவாகும் தொடர் வினைத்தொடர் எனப்படும். வினைத்தொடரில் வினைக்கு முன்னர் வேற்றுமைத்தொடர், பின்னுருபுத்தொடர், வினையடை, வினையெச்சத்தொடர் என்பன வரலாம்.

எ.கா.

விழுந்தேன், செய்தேன், அங்கு வந்தேன்

நன்றாக விளையாடினேன், வந்து பார்த்தேன்

போன்றவை வினைத்தொடர்களாகும்.

7.5.1.2 வாக்கிய வகைகள்

வாக்கியங்களைப் பல்வேறு அடிப்படையில் பல வகைகளாகப் பிரிப்பர் இலக்கணயலார். அமைப்பினை அடிப்படையாக வைத்துத் தமிழ் வாக்கியங்களை வினைப்பயனிலை வாக்கியங்கள் என்றும், பெயர்ப் பயனிலை வாக்கியங்கள் என்றும் பிரிப்பர். இதுபோன்று வாக்கியங்களின் சேர்க்கையை அடிப்படையாகக் வைத்து தனி வாக்கியம் அல்லது எளிய வாக்கியம் (Simple sentence) என்றும், கலவை வாக்கியம் (Complex sentence) என்றும் என்றும் பிரிப்பர். வெவ்வேறு வினைகளின் பண்பையும் அதனால் அடையும் மாற்றங்களையும் வைத்து செய்வினை வாக்கியம் (active voice) என்றும், செயப்பாட்டு வினை வாக்கியம் (passive sentence)என்றும் பிரிப்பது உண்டு. இவை மட்டுமின்றி வேறொரு பண்பினை ஒட்டி இலக்கணங்கள் வாக்கியங்களை வினா வாக்கியம் (interrogative sentence) என்றும், செய்தி வாக்கியம் (affirmative sentence) என்றும் பிரிப்பர். இவை மட்டுமல்லாமல் இலக்கணவியலார் ஏவல் வாக்கியம், செய்தி வாக்கியம் என்றும் பிரிப்பர். தமிழில் காணப்படும் வாக்கியங்களை 1. வினைப் பயனிலை வாக்கியம், (கண்ணன் வந்தான்), 2. பெயர்ப் பயனிலை வாக்கியம் (இராமன் ஒரு ஆசிரியர்), 3. பெயரடை பயனிலை வாக்கியம் (இராமன் நல்லவன்), 4. உடைமைப் பயனிலை

வாக்கியம் (இராமன் என்னுடையவன், புத்தகம் என்னுடையது) என நான்காகப் பிரித்து இனம் காண்பர் சிலர் (கோதண்டராமன்,1997) .

தொடரியலில் சில முக்கியமான அமைப்புகளை மட்டுமே குறித்துச் சொல்லுகின்ற நிலையில் இதன் கண் ஒருசில அமைப்புகள் மட்டுமே எடுத்துக் கூறப்படுகின்றன.

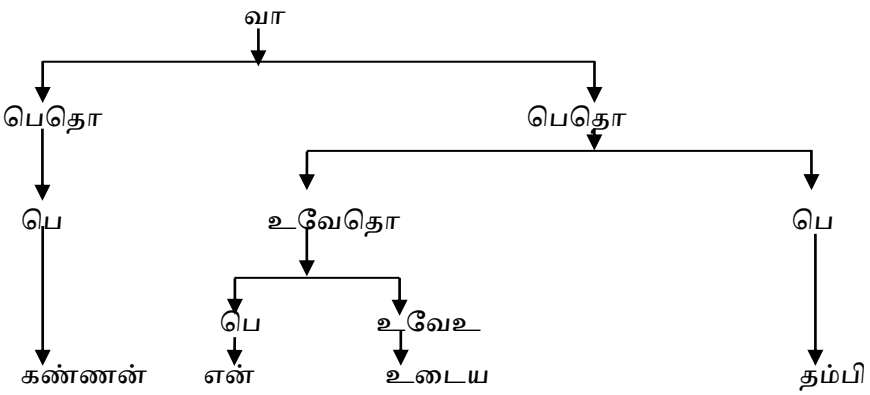
7.5.1.2..1 அடிப்படை வாக்கிய வகைகள்

தமிழில் காணப்படும் அடிப்படையான வாக்கியங்கள் 1. பெயர்ப் பயனிலை வாக்கியம் என்றும், 2. வினைப் பயனிலை வாக்கியம் என்றும் இரு வகைகளாகப் பிரிக்கலாம். இவ்விரண்டு வாக்கிய வகைகளிலும் பெயர்த்தொடர் எழுவாய்களாக வரும் இந்நிலையில் இவற்றை பின்வரும் அமைப்பைக் கொண்டுள்ளவைகளாகப் பிரிக்கலாம்.

1.பெயர்த்தொடர் + பெயர்த்தொடர் (NP + NP)

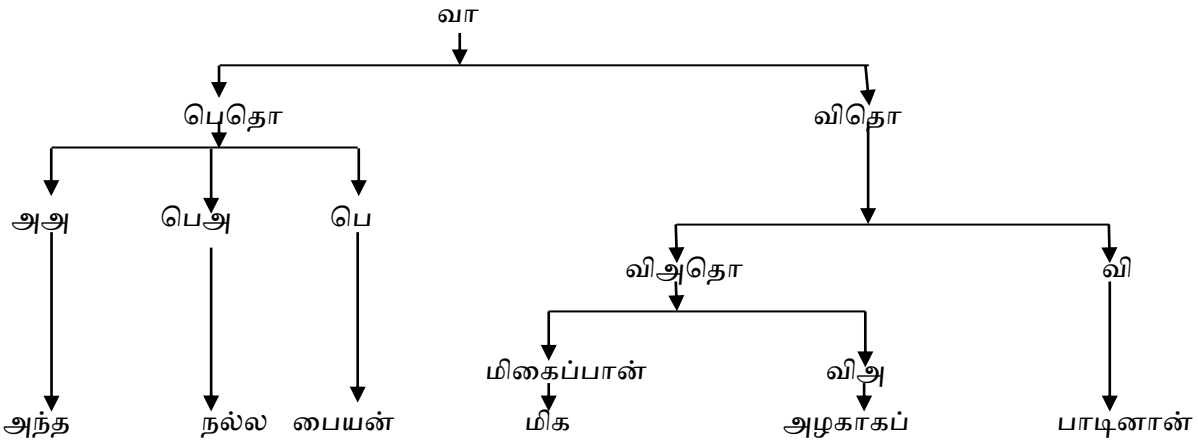
2.பெயர்த்தொடர் + வினைத்தொடர் (NP + VP)

ஒரு பெயர்ப் பயனிலைத் தொடர் வாக்கியம் எவ்வாறு உள்ளது என்பது கீழே உள்ள கிளைப்படத்தில் காட்டப்பட்டுள்ளது.



மேலே 'கண்ணன் என்னுடைய தம்பி'என்ற பெயர்ப் பயனிலை வாக்கியத்தின் அமைப்பு காட்டப்பட்டுள்ளது.

வினைப் பயனிலை வாக்கியம் என்பது ஒரு வினையையும் வினையடைத் தொடரையும், மிக என்ற மிகைப்பாணையும் கொண்டு வந்துள்ளது.



7.5.1.2..2 வினா வாக்கியங்கள்

வினா வாக்கியம் என்பது தெரியாத ஒன்றை மற்றொருவரிடம் கேட்டுத் தெரிந்து கொள்ளுதற்காகப் பயன்படுத்தப்படும் வாக்கியம் ஆகும். வினாவில் அடிவினா என்றும், ஒட்டுவினா என்றும் இருவகை உள்ளன. அடிவினா யார், என்ன போன்ற வினாச் சொற்களையும் ஒட்டுவினா ஆ என்ற ஒட்டையும் (எ.கா. அவனா? இவனா?) கொண்டிருக்கும்.

வினா வாக்கியங்களும் அவற்றின் விடையாக வரும் செய்தி வாக்கியங்களும் ஏகதேசம் ஒரே அமைப்பைக் கொண்டு விளங்கும்.

ஆம் இராமன் வந்தான்.

அவன் நேற்று வந்தான்.

என்ற செய்தி வாக்கியங்களும்,

இராமன் வந்தான். இராமன் வந்தானா?

அவன் நேற்றா வந்தான்?

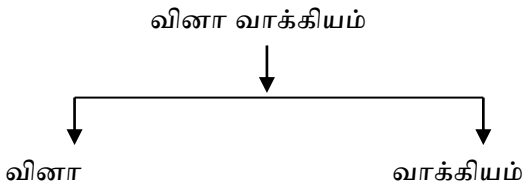
என்ற ஒட்டு வாக்கியங்களும் ஏகதேசம் ஒரே அமைப்பைக் கொண்டவை. ஒன்றில் வினா ஒட்டு உள்ளது. செய்தி வாக்கியத்தில் அது இல்லை. இதே நிலைதான் அடிவினா வரும் வாக்கியங்களிலும் காணப்படுகின்றது.

இராமன் வந்தான். இராமன் என்று வந்தான்?

இராமன் ஒரு ஆசிரியர். இராமன் யார்?

அவன் நன்கு செய்தான் அவன் எப்படிச் செய்தான்?

போன்ற இணை வாக்கியங்களும் ஒரே மாதிரியான அமைப்பையே கொண்டுள்ளன. ஒன்றில் செய்தி (நேற்று) மற்றொன்றில் வினா (என்று) இதனால் மொழியியலாளர்மள் வினா வாக்கியங்களையும் அதை ஒட்டி வரும் செய்தி வாக்கியங்களையும் ஒரே அமைப்பின் கீழ்க் கொண்டு வந்து வினா வாக்கியத்தைப் பிரித்துக்காட்டும் நிலையில் 'வினா' என்ற முன்னிலையை இணைத்து வேறுபடுத்துவர். இதனைக் கீழ வருமாறு காட்டலாம்.



ஒட்டு வினாவும், அடிவினாவும் ஒன்று வந்த இடத்தில் மற்றொன்று வருவதில்லை. இதனைக் கீழ்வருமாறு காட்டலாம்.

வினா-ஒட்டுவினா, அடிவினா

ஒட்டுவினாப் பொருள்

ஒட்டுவினா ஆம் அல்லது இல்லை என்ற பதில்களை எதிர்பார்ப்பது. முதலில் உள்ளது கேட்போரின் உள்ளத்தில் உள்ளதை உறுதி செய்வது . இரண்டாவது உள்ளதை மறுப்பது.

இராமன் வந்தான்.

இராமன் வந்தானா?

எனத் தமிழில் வருவது காணத்தக்கது. முதல் வினாவின் விடையாக ஆம் (இராமன் வந்தான்) என்றோ இல்லை (இராமன் வரவில்லை) என உறுதி செய்தோ அல்லது மறுத்தோ விடை அமையலாம். இதுபோன்றே அடுத்த வினாவிற்கும் பதில்கள் அமையலாம்.

தமிழில் ஒட்டுவினா, வினா-ஆ என்ற உருபனால் காட்டப்படுகின்றது. பொதுவாக இது வாக்கியங்களில் ஒரே ஒரு இடத்தில் வரும் பல்வேறு இலக்கணக் கூறுகளில் ஏதாவது ஒன்றின் பின்னால் வந்து ஒட்டு வினாப்பொருளைக் காட்டும். இவ்வாறு ஏதாவது ஒன்றின் பின்னால் வருவதை வினாவின் வீச்சு (ளஉடிப்ந) என்பர் மொழியியலாளர்கள்.

இராமன் + வினா + வந்தான் >இராமன் வந்தான்.

இராமன் வந்தான் + ஆ >இராமன் வந்தானா?

என்று வருவது குறிப்பிடத்தக்கது.

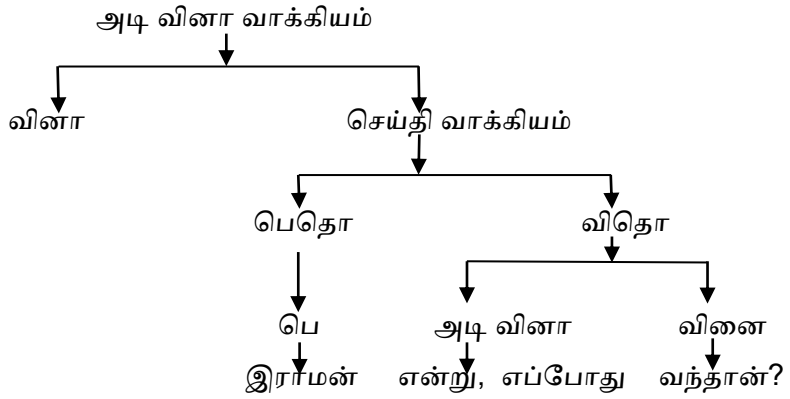
அடி வினா

அடி வினா வாக்கியங்கள் 1.யார், எவர், எவன், என்ன, எது, எவை போன்ற வினாப் பெயர்களையும், 2.எந்த, எத்தனையாவது, எத்தனை, எவ்வளவு, எப்பேர்பட்ட அல்லது எப்படி உள்ள (எத்தகைய, எம்மாதிரி போன்ற) என்ன போன்ற வினாப் பெயரடைகளையும், 3.எத்தனையாவது, எங்கு, எப்படி, எப்போது போன்ற வினா வினையடைகளையும் உள்ளடக்கியவை. இவற்றுள் சில ஒன்றுக்கு மேற்பட்ட நிலைகளிலும் (என்ன, எப்பொழுது) வரும்.

இராமன் நேற்று வந்தான்.

இராமன் என்று வந்தான்?

போன்ற வாக்கியங்கள் ஒரே அமைப்பைக் கொண்டுள்ள நிலையில் 'நேற்று' 'என்று' போன்றவை ஒரே அமைப்பு கொண்ட வாக்கியங்களில் ஒரே இடத்தில் வருபவை ஆகும். எனவே அடி வினா வாக்கியங்களின் அமைப்பைக் கீழ்வருமாறு காட்டலாம்.



என்ற ஒரே அமைப்பைக் கொண்டுள்ள நிலையில் அடி வினா என்ற புனை நிலையில் (hypothetical) கொள்ளப்பட்டது கெட அடி வினா வாக்கியம் உருவாக்கப்பெறுகிறது.

7.5.1.2.3 செய்வினை, செயப்பாட்டு வினை வாக்கியங்கள்

வினாப் பாகுப்பாடு பற்றிக் கூறும்போது வினைகள் செய்வினை என்றும் செயப்பாட்டு வினை என்றும் பகுக்கப்படும் என்று கூறப்பட்டது. செய்வினைப்யன் செயவென் வாய்பாட்டு வினஇயெச்சத்தின் பின்னர் படு அல்லது பெறு என்ற துணை வினைகளை இணைக்கும்போது உண்டாகும் வினைதான் செயப்பாட்டு வினை எனவும் கூறப்படுகிறது. அவ்வாறு உள்ள

செயப்பாட்டு வினையைப் பயனிலையாகக் கொண்ட வாக்கியங்கள் செயப்பாட்டுவினை வாக்கியங்கள் ஆகும்.

செயப்படுபொருளை எழுவாயாக ஆக்கியும் செய்பவனை அல்லது மருத்தாவை ஆல் வேற்றுமையாக ஆக்கியும் கூறப்படுவது செயப்பாட்டு வினை வாக்கியம் (Passive Sentence).

இராஜா இராமனை அடித்தான்

என்பது செய்வினை வாக்கியம். இங்கு செய்பவன் எழுவாயாகவும், செய்யப்படுபவன் ஐ வேற்றுமையாகவும் உள்ளது. இதனை உரிய இலக்கண மாற்றத்தால் மாற்றி

இராமன் இராஜாவால் அடிக்கப்பட்டான்

என்ற நிலையில் உருவாக்கம் பெறுகின்றது. இங்குச் செய்வினை வாக்கியம் அடிப்படை வாக்கியமாகவும் (புகைநிலை), செயப்பாட்டு வினை வாக்கியம் வருவிக்கப்பட்ட வாக்கியமாகவும் (புறநிலை) காணப்படுகின்றன. இவற்றிற்கு இடையே காணப்படும் மாற்றத்ததைக் கீழ்வருமாறு காட்டலாம்.

பெ.தொ.1+பெ.தொ.2_ஐ+வினை+காலம்+விகுதிகள்

செ.பா

பெ.தொ.2+பெ.தொ.1-ஆல்

வினை. செய.எ.எச்சம்+பகு+காலம்+விகுதிகள்

என்ற நிலையில் உருவாக்கம் பெறுகின்றது.

7.5.1.2.4 கலப்பு வாக்கியம்

ஒரு வாக்கியத்தின் உள்ளே இன்னொரு வாக்கியத்தை முன்னதன் பகுதியாக அல்லது உறுப்பு வாக்கியமாக இணைக்கும்போது உருவாகும் வாக்கியம் கலப்பு வாக்கியம் எனப்படும். எந்த வாக்கியம் தலைமை வாக்கியம் (Matrix sentence) என்றும், இணைக்கப்படுகின்ற வாக்கியம் உறுப்பு வாக்கியம் (Constituent sentence) என்றும் அழைக்கப்படும். எனவே கலப்பு வாக்கியம் என்பது ஒரு தலைமை வாக்கியத்தையும் ஒன்றோ அல்லது ஒன்றுக்கு மேற்பட்ட உறுப்பு வாக்கியத்தையும் கொண்டிருக்கும் வாக்கியம் ஆகும்.

நான் இராமன் வந்தபோது போனேன்

நான் அங்கு வந்த பையனைப் பார்த்தேன்

போன்றவை கலப்பு வாக்கியங்கள்.

கால வினையெச்சத் தொடர் வாக்கியம்

இராமன் வந்தபோது நான் போனேன்

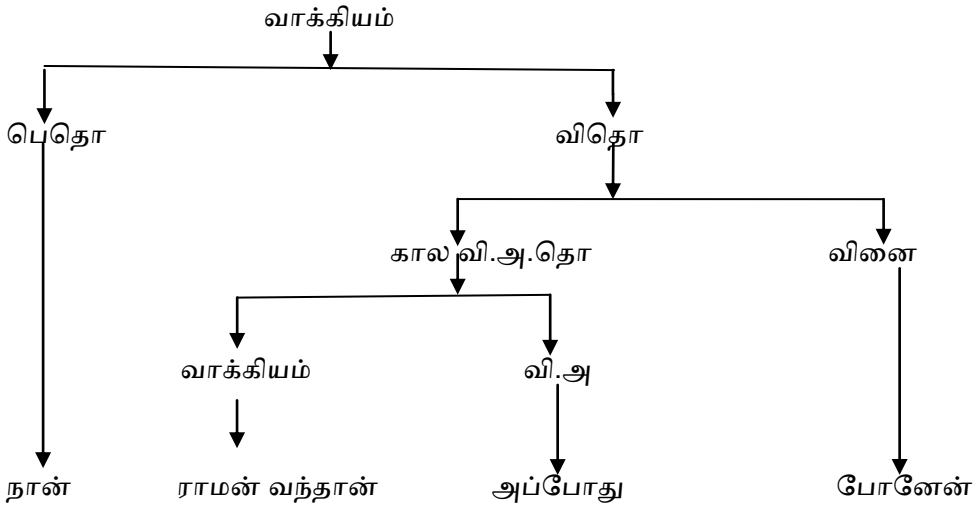
என்பது ஒரு கலப்பு வாக்கியம். இவ்வாக்கியம்

இராமன் வந்தான், அப்போது நான் போனேன்

என்ற இரு வாக்கியங்கள் ஒன்றில் மற்றொன்று உறுப்பு வாக்கியமாக இணைவதால் ஏற்படும் ஒரு வாக்கியம் ஆகும். இதில் இரண்டாவது வாக்கியம் முக்கிய அல்லது தலைமை வாக்கியமாகவும் முதல் வாக்கியம் அதில் கலந்த உறுப்பு வாக்கியமாகவும் உள்ளன. *அப்போது* என்ற வினையடையை விளக்கும் நிலையில் அமைந்துள்ளது *இராமன் வந்தான்* என்ற வாக்கியம்.

புதைநிலை (Deep structure)

இவ்வாக்கியத்தின் புதைநிலையைக் கீழ்வருமாறு காட்டலாம்.



பல்வேறு மாற்றங்கள் (Transformations)

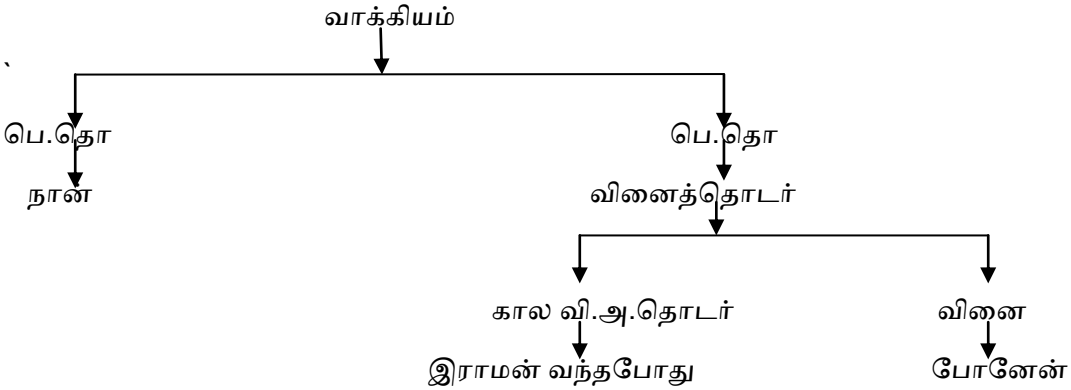
இப்புதைநிலையில் பல்வேறு மாற்றங்கள் நிகழ்கின்றன.

1. வந்தான் என்பது விகுதி. இதுவே பெயரெச்ச விகுதியாகிறது. இதனைப் பெயரெச்ச மாற்றம் (relativisation) எனக் கூறலாம்.

2. அப்போது என்பதில் உள்ள அ என்பது கெடுகின்றது. இதனை அகரச் சுட்டு கெடுதல் எனக் கூறலாம்.

புறநிலை (surface structure)

மேலே காட்டிய மாற்றங்கள் நிகழும்போது *நான் இராமன் வந்தபோது* போனேன் என்ற புறநிலை வாக்கியம் உருவாகின்றது.



என இதனைக் காட்டலாம்.

பிற கலப்பு வாக்கியங்கள்

பலவிதமான கால எச்சத் தொடர்களும், வேற்றுமைத் தொடர்களும், பின்னுருபுத் தொடர்களும், வினையெச்சத் தொடர்களும், பெயரெச்சத் தொடர்களும் தலைமை வாக்கியத்தின் உறுப்புகளாக வந்து கலப்பு வாக்கியங்களை உருவாக்கும்.

உடனிகழ்ச்சி கலப்பு வாக்கியம்

மணி பாட நான் ஆடினேன்

என்ற வாக்கியத்தில் இரு நிகழ்வுகளைக் குறிக்கும் வாக்கியங்கள் ஒன்றிற்குள் ஒன்றாக உள்ளடக்கப்பட்டு கலப்பு வாக்கியமாக வெளிப்படுகின்றன.

வேற்றுமைத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் வேற்றுமைத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

நான் கண்ணன் வந்ததைக் கண்டேன்

பின்னுருபுத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் பின்னுருபுத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

நான் வருவதற்கு முன்னர் அவள் வந்து விட்டாள்.

வினையெச்சத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் வினையெச்சத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

அவள் சாப்பிட்டுவிட்டு தூங்கினாள்

பெயரெச்சத் தொடர் வாக்கியம்

இதில் ஒரு வாக்கியம் பெயரெச்சத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

நேற்று ஊருக்குப்போன கண்ணன் இன்று பள்ளிக்கு வரவில்லை.

7.5.1.2.5 கூட்டு வாக்கியம்

இரண்டு வாக்கியங்கள் தலைமை வாக்கியம், உறுப்பி வாக்கியம் என்றில்லாமல் ஒத்த நிலையில் இணையும்போது உருவாகும் வாக்கியம் கூட்டு வாக்கியம் எனப்படும்.

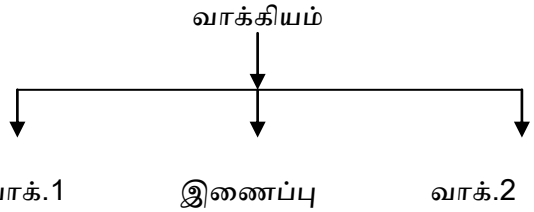
கண்ணன் வந்தான்.

இராமன் வந்தான்.

என்ற இரு வாக்கியங்களும் ஒத்தநிலையில் இணையும்போது

கண்ணனும் இராமனும் வந்தார்கள்.

என்ற கூட்டு வாக்கியம் உருவாகின்றது. இதில் *கண்ணனும் இராமனும்* என்ற பெயர்த்தொடர் எழுவாயாகவும், *வந்தார்கள்* என்பது பயனிலைத் தொடராகவும் உள்ளன. இரண்டு வாக்கியங்களும் கூட்டு வாக்கியத்தில் ஒத்த நிலையில் காணப்படுகின்றன. ஒன்றின் உறுப்பாக மற்றொன்று இல்லை. இத்தகைய இணைப்பு அமைப்பைக் கீழ்க்கண்டவாறு காட்டலாம்.



கூட்டு வாக்கியங்களின் வகைப்பாடு

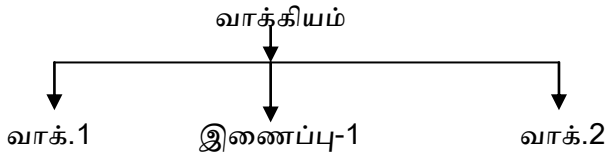
இரு வாக்கியங்கள் இணைந்து உருவாகும் கூட்டு வாக்கியங்கள் (1) முழுநிலை இணைப்பு வாக்கியங்கள் (conjoined sentences) என்றும், (2) பிரிநிலை இணைப்பு வாக்கியங்கள் (disjunctive) என்றும் இரண்டு வகைப்படும்.

முழுநிலை இணைப்பு

முழுநிலை இணைப்பு என்பது ஒன்றுக்கு மேற்பட்ட வாக்கியங்கள் முழு நிலையில் இணைக்கப்படுவது. அதாவது அதுவா இதுவா என்ற நிலையில் இணையாது அதுவும், இதுவும் என்ற கூட்டுநிலையில் இணைக்கப்படுவது (conjunctive).

கண்ணனும் இராமனும் வந்தார்கள்.

இதனைப் பின்வருமாறு காட்டலாம்.



பிரிநிலை இணைப்பு வாக்கியம்

இரு வாக்கியங்களை வேறுபடுத்தி அல்லது பிரித்து இணைக்கும் இணைப்பு பிரிநிலை இணைப்பு (disjunctive) எனப்படும். அதாவது 'அதுவா இல்லை இதுவா' என்ற நிலையில் இணைக்கப்படுவது

கண்ணன் வருவான் அல்லது இராமன் வருவான்.

கண்ணன் அல்லது இராமன் வருவான்.

கண்ணனோ அல்லது இராமனோ வருவார்கள்

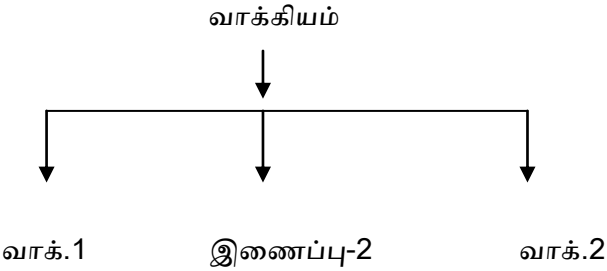
போன்ற வாக்கியங்கள் இப்பிரிவைச் சாரும்.

கண்ணன் வருவான் + இணைப்பு2 + இராமன் வருவான்

என்ற இரு வாக்கியங்களிலிருந்து உருவாவதுதான்.

கண்ணன் வருவான் அல்லது இராமன் வருவான்

போன்ற வாக்கியங்கள் கீழ்வருமாறு காட்டப்படுகிறது.



தமிழ்த்தொடரியல் அமைப்பில் ஒரு பெயர்த் தொடரும் ஒரு வினைத் தொடரும் இணைந்தோ இரு பெயர்த் தொடர்கள் இணைந்தோ ஒரு வாக்கியத்தை உருவாக்கும் வாக்கியங்கள் அவற்றின்

அமைப்பின் அடிப்படையில் தனிநிலை வாக்கியம், கூட்டு வாக்கியம் மற்றும் கலவை வாக்கியம் என வேறுபடும். தனிநிலை வாக்கியத்தில் ஒரு வினைத்தொடர் காணப்படும். கூட்டு வாக்கியத்தில் ஒன்றிற்கும் மேற்பட்ட வினைத்தொடர்களோ, பெயர்த்தொடர்களோ இணைப்பான்களால் இணைக்கப்பட்டிருக்கும். கூட்டுநிலை வாக்கியத்தில் ஒன்றுக்கும் மேற்பட்ட வினைத்தொடர்கள் (அதாவது எச்சத் தொடர்கள்) வரும். பொருண்மை அடிப்படையில் வாக்கியங்களைக் கூற்று வாக்கியங்கள், கட்டளை வாக்கியங்கள், வினா வாக்கியங்கள், செய்வினை வாக்கியங்கள், செயப்பாட்டு வினை வாக்கியங்கள் எனப் பகுக்கலாம். ஒரு வாக்கியம் மற்றொரு வாக்கியத்தில் உட்படையாக வருவதன் காரணமாக வாக்கியங்கள் கலவைத்தன்மை பெறுகின்றன. பெயரெச்சத் தொடர்கள் பெயர்த் தொடருக்குள்ளும் வினையெச்சத் தொடர்கள் வினைத் தொடர்க்குள்ளும் உட்படையாக வரும்.

7.5.2 முந்தைய கணினித் தொடரியல் ஆய்வுகள்

முந்தைய பகுத்துக்குறித்தல் அணுகுமுறைகள் பெரும்பாலும் அகராதியால் இயக்கப்பட்டன. இதில் அகராதி உருபனியல் தரவையும், பொருண்மையியல் வர்ணனைகளையும், தொடரியல் அறிவையும் கொண்டிருக்கும். பின்னர் இலக்கண வடிவ வாதங்களின் தொடக்கம் பல வேறுபட்ட கொள்கைகள் அடிப்படையிலான பகுத்துக் குறிப்பான்களைத் தந்தது. காஷ்கெட் (Kashket,1986) என்பவர் வார்ல்பிரி என்ற மொழிக்கு சாம்ஸ்கிலின் (Chomsky) ஆளகை-கட்டுறவுக் கோட்பாடு (government and binding) அடிப்படையில் பகுத்துக்குறிப்பானை உருவாக்கினார். இவ்வணுகுமுறையில் தொடர்கள், தேர்வுக் கொள்கைகள், வேற்றுமை அடையாளப்படுத்துகை, வேற்றுமை தருகை, பங்கெடுப்பாளர் தொடர்பு படுத்துகை என்பனவற்றின் வழி உருவாக்கப்பட்டன. பங்கெடுப்பாளர் தகவல் அகராதியில் சேகரிக்கப்பட்டன. சென்குப்தா (Sengupta, et al., 1997) போன்றோர் கப்லன் மற்றும் பிரஸ்னைன் என்போரின் சொல் செயல்பாட்டு இலக்கணத்தை (Lexical Fuhnctional Grammar)வங்காள மொழிக்குத் தகுந்தவாறு மாற்றி ஒரு பகுத்துக்குறிப்பானை உருவாக்கினார். சொல் செயல்பாட்டு இலக்கணம் ஆங்கிலம் போன்ற இடம்சார் மொழிகளுக்கு (Positional languages) வலுவான வடிவவாதம் ஆகம். இது உறுப்பமைபபையும் இலக்கண உறவுகளையும் கண்டுகொள்கிறது. செங்குப்தாவின் பகுத்துக் குறிப்பான் செய்வினைப்பாடாக உள்ள எளிய வாக்கியங்களுக்கு நல்ல விளவைத் தந்தது. வினிவார்ட்டர் Winiwarter,1995)ஒன்றிணைப்பு

இலக்கணத்தை (unofication grammar) விரிவாக்கி ஜோமன் மொழிக்குப் பயன்படுத்தினார். சினிவார்ட்டரின் ஒழுங்கமைப்பில் இலக்கண விதிகள் அருவமாக்கலின் பொருத்தமான மட்டத்தில் இலக்கண விதிக்ள அதராதிக்கு பங்கெடுப்பாளராகச் சேர்ககப்ப்பட்டன. பெக்கா போன்றோர் (Becker et.al, 1994) V_TAG என்ற கிளை இணைப்பு இலக்கணத்திலிருந்து (TREE adjoining) மாறுபட்ட ஒன்றான பலபெயர் கால பகுத்துக்குறித்தல் வழிமுறை வரைவை உருவாக்கினார் இப்பகுத்துக்குறிப்பானை கிளை இணைப்பு இலக்கண வடிவவாதம் முற்றிலும் உருவாக்கப்பட்ட மொழிகளுக்கு மட்டுமே பயன்படுத்த இயலும்.

பல பகுத்துக்குறிப்பான்கள் சார்புக் கொள்கைகள் (dependency principles) அடிப்படையில் உருவாக்ககப்பட்டன. கவிங்டன் (Covington, 2003) உரு சுற்று வரைபடத்தால் இயக்கப்பட்ட சொற்களுக்கு இடையே உள்ள சார்புகளை உருவாக்கி வாக்கியங்களுக்கு அமைப்பைத் தரும் ஒரு பகுத்துக்குறிப்பானை உருவாக்கினார். இதில் ஒவ்வொரு வில்லும் சொற்களின் சார்புநிலை உறவை உருப்படுத்தம் செய்யும். காரக உறவுகள் வேற்றுமைக் குறிகளுடன் பொருத்தப்படும் பாணினி இலக்கண அடிப்படையில் பாரதி போன்றோர் Gharti, et, al,1993, 2003) கட்டுப்பாடு அடிப்படையிலான பகுத்துக் குறிப்பான்களின் ஒரு குழுமத்தை உருவாக்கினார். பாரதி ஒழுங்குமுறை பகுத்துக்குறிக்கும் சிக்கலை முழுஎண் வழியமைப்புச் சிக்கலாக மொழி பெயர்த்துத் தீர்வு காண்கிறது. மூர்த்தி (Murthy,1997) இந்திய மொழிகளுக்கு உரக எச்சத்தொடர் அமைப்பு இலக்கணம் (Universal Clause Structure Grammar) என்ற இலக்கண வடிவ வாதத்தை உருவாக்கினார். இவ்வடிவவாதத்தில் மூன்று மட்ட பகுப்பாய்வு பயன்படுத்தப்படுகிறது. தொடர் அடையாளம் காணல், எச்சத்தொடர் அமைப்பு நிர்ணயம் மற்றும் பங்களிப்பு தருதல். ஒவ்வொரு மட்டத்திற்கும் அதன் பகுப்பாய்வுக்குத் தொடர்புள்ள விதிகள் இருக்கின்றன. உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தில் இத்தொகுதியாக்கம் (modularization) இதைப் பிற மொழிகளுக்கு எளிதாகப் பயன்படுத்த செய்தது. சொல்நிரல் சிக்கலைக் கையாளுவதற்கான ஒரு தீர்மாங்களில் ஒன்று வரிசை மாற்று இலக்கணமாகும் (Permutational Grammar (PG)). (Geg. Ikifssib, et, al., 2001) வரிசை மாற்று இலக்கணம் அடிப்படை தொடரமைப்பு விதிகளை அவற்றின் செயல்பாட்டு உருப்படுத்தங்களுடன் குறிப்பிட்டு பின்னர் ஒரே பொருண்மையைத் தரும் பிற எல்லாத் தொடர்ச்சிகளையும் பெறவேண்டி உறுப்புகளை வரிசைமாற்றம் செய்யும். குமாரசண்முகம் மேலே குறிப்பிட்ட வேறுபட்ட அணுகுமுறைகளின் ஒரு ஒன்றிணைப்பைப் பயன்படுத்துகின்றார். அவர்

தொடர்களையும் வாக்கியங்களையும் எச்சத்தொடர்களையும் கண்டுபிடிக்க மூர்த்தியின் உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தைப் போன்று நுட்பங்களைப் (Techniques) பயனபடுத்துகின்றார். சுதந்திர சொல்நிரல் சிக்கலைக் கையாள வரிசைமாற்று இலக்கணத்தைப் போன்று உறுப்புகளை வரிசை மாற்றுவதில்லை. அதற்குப் பதிலாக வாக்கியங்களின் உறுப்புகளை ஒரு குறிப்பிட்ட நிரலில் வரிசைமாற்றம் செய்கிறார். பாரதியின் ஒழுங்கமைப்பில் (Bharati, 2000) உள்ளது போன்று வாக்கியத்தின் எழுவாயையும் செயப்படுபொருளையும் கண்டுபிடிக்க வேற்றுமை உருபுகளைப் பயன்படுத்துகிறார். தமிழின் தொடரியலை விளக்க பொதுவான தொடரமைப்பு இலக்கணததைப் பயன்படுத்துகிறார்.

தமிழ் வாக்கியங்களைப் பகுத்துக் குறிப்பதற்கு வேண்டி ஒரு கணினியியல்சார் சட்டகத்தை உருவாக்கி பயன்படுத்துவதாகும். இதற்கு வேண்டி தமிழ்த்தொடரியலை உருப்படுத்தம் செய்ய தொடரமைப்பு விதிகள் உருவாக்கப்பட்டுள்ளன. நாற்பது மிகக் கூடதலாக நிகழும் வினைகளுக்குப் பங்கெடுப்பாளர் அமைப்பு கண்டுபிடிக்கப் பட்டுள்ளது. பகுத்துக்குறிக்கப்பட்ட வாக்கியங்களின் தொடரியல் அமைப்புகள் பகுத்துக்குறிக்கப்பட்ட கிளைகளில் உருப்படுத்தம் செய்யப்பட்டுளளன. இவ்வடிவில் அவற்றை இயந்திர மொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுப்பு இவற்றை உள்ளடக்கிய பயன்பாடுகளில் நேரடியாகப் பயன்படுத்த இயலும்.

உள்ளீட்டுத் தமிழ் வாக்கியத்தை முக்கியப் பகுத்துக்குறிப்பான் பகுத்துக் குறிப்பதற்கு முன் அது பகுப்பாய்வின் பல மட்டங்களைக் கடந்துசெல்ல வேண்டும். கீழே தந்துள்ள படத்தில் செவ்வகங்கள் பகுப்பாய்வு செய்யும் பகுப்பாய்வுத் தொகுதிகளையும் (analyzer modules) வட்டங்கள் அவை பயன்படுத்தும் பொருத்தமான மூலப்பொருள்களையும் உருப்படுத்தம் செய்யும். தமிழுக்கு வேண்டி சில பெயர்த்தொடர் பகுப்பான்கள் உருவாக்கப்பட்டுள்ளன. அவை கீழே விளக்கப்பட்டுள்ளன.

7.5.2.1 AUKBCRC -இன் தமிழுக்கான பெயர்த்தொடர் பகுப்பான்

இவ்வணுகுமுறை விதி அடிப்படையிலானது. இந்நெறிமுறையில் ஒரு தரவுத்தொகுதி எடுக்கப்பட்டு இரண்டோ அதற்கு அதிகமான குமங்களாகப் பிரிக்கப்படும். இப்பிரிக்கப்பட்ட குழுமத்தில் ஒன்று பயிற்சித் தரவாகப் பயன்படுத்தப்படும். பயிற்சித் தரவுக் குழுமம் எடுக்கப்பட்டு பெயர்த்தொடருக்காக மனித முயற்சியால் பகுக்கப்படும். இவ்வாறு ஒரு வாக்கியத்தில் உள்ள பெயர்த் தொடர்களைப் பிரிப்பதற்குப் பயன்படுத்தப்படும் விதிகள் உருவாக்கப்படும். இவ்விதிகள்

தொடர் பகுப்பிற்கு அடிப்படையாக அமையும். தொடர் பகுப்பான் வழியமைப்பு இவ்விதிகளைப் பயன்படுத்தி பரிசோதனைத் தரவை பகுக்கும். இவ்விதிகளின் பயன்பாட்டுப் பரப்பு இத்தரவுக்குழுமத்தில் பரிசோதிக்கப்படும். துல்லியமும் மீள அழைத்தலும் இதற்குக் கணிக்கப்படும். விளைவு ஒழுங்குமுறையின் பயன்பாட்டுப் பரப்பை விருத்தி செய்ய கூடதல் விதிகள் தேவையா என்று பரிசோதிக்கப்படும். பின்னர் ஒழுங்குமுறை வேறுபட்ட பிற பயன்பாடுகளுக்குப் பரிசோதிக்கப்படும்.

7.5.2.2 RCILTS –இன் வானவில்

வானவில் இந்திய மொழிகளின் தொழில்நுட்ப முன்னேற்றம் என்ற திட்டத்தின் கீழ் தமிழுக்காகச் செய்யப்பட்ட இயற்கைமொழி ஆய்வுகளில் ஒன்றாகும். இது ஒரு தமிழ் வாக்கியத்தின் தொடரியல் உறுப்புகளை அடையாளம் காண்கிறது. இது பட்டியல் வடிவில் பகுப்புக் கிளைப் படத்தை (parse tree) வெளிப்யிடுகின்றது. இது தனிநிலை வாக்கியங்களைய கூட்டு வாக்கியங்களையும் கையாளுகின்றது. தனிநிலை வாக்கியங்கள் ஒரு வினையையும் பல பெயர்த் தொடர்களையும், தனிநிலை வினையடைகளையும், பெயரடைகளையும் கொண்டிருக்கும். பல எச்சத் தொடர்கள் உள்ள வாக்கியங்களின் தேர்வுகளில் வானவில் துப்புச்சொற்கள் மற்றும் தொடர்கள் அடிப்படையில் எச்சத் தொடர்களைத் தொடர் குழுக்களாகக் குழுமும். இது தொடரியல் அமைப்பு இலக்கணத்தை உருவாக்கும். வானவில் சொல்நிரலைக் கையாள முன்நோக்குதலைப் பயன்படுத்தும்.

பகுத்துக்குறித்தல் என்பது ஒரு வாக்கியத்தை அறிந்துகொண்டு அதற்குத் தொடரியல் அமைப்பைத் தருவதாகும். இது தொடரியல் ஆய்வில் மிகப்பெரிய சிக்கலாகும். இதன் காரணமாக இது இயற்கை மொழி ஆய்வின் ஒரு பெரும்பகுதியாக அமைகிறது.

நிர்ணயிக்கப்பட்ட பகுத்துக்குறிப்பானின் உள்ளீடு (input) அடையாளப்படுத்தப்பட்ட சொற்களின் நிரலான தொடர்ச்சியும் தொடரமைப்பு விதிகளின் ஒரு குழுமமும் ஆகும். இதன் வெளிப்யீடு (out put) வாக்கியத்தின் பகுத்துக்குறிக்கப்பட்ட அமைப்பாகும். ஒரு வாக்கியத்தின் பகுத்துக் குறிக்கப்பட்ட அமைப்பு அதன் எழுவாய், செயப்படுபொருள் மற்றும் பயனிலைகளை அறிந்துகொள்ள உதவுகிறது. மேலும், பகுத்துக் குறிக்கப்பட்ட கிளை அமைப்பு பிற தொடர்களையும் அடை செய்யும் தொடர்களையும் அடையாளம் காணப்படுகிறது.

ஒரு மொழிக்கு ஒரு மாதிரி பகுத்துக்குறிப்பான் ஒரு அடையாளப்படுத்தியைப் பயன்படுத்தி வாக்கியங்களின் சொல் வகையை அடையாளப் படுத்துகிறது. பின்னர் பகுத்துக்குறித்தும் வழிமுறை வரைவு (algorithm) வாக்கியத்தைப் பகுத்துக்குறித்து அம்மொழி இலக்கணத்திற்குத் தரப்பட்ட தொடரமைப்பு விதிகிளின் அடிப்படையில் பொருத்தமான அமைப்பைத் தரும். பகத்துக்குறித்தலின் போது வழிமுறை வரைவு உறுப்புகளுக்கு இடையேயுள்ள இலக்கண உடன்பாட்டையும் முக்கிய வினைக்கம் செயப்படுபொருள் தொடர்களுக்கும் இடையே உள்ள சார்பு உறவையும் பரிசோதிக்கும்.

7.5.2.3. குமாரசண்முகத்தின் தொடரியல் பகுப்பாய்வு

குமார சண்முகம் தமது எம்.எஸ். பட்டப்படிப்புக்காக தமிழில் தொடரியல் பகுப்பாய்வு என்ற ஆய்வேட்டைச் சமர்ப்பித்தார். அதன் சுருக்கம் இங்கு தரப்பட்டுள்ளது.

7.5.2.3.1. வாக்கியங்களின் இலக்கணப் பண்புக்கூறுகள்

வாக்கியத்தில் உள்ள சொற்களின் சொல் வகைப்பாடும் பண்புக்கூறு தகவலும் வாக்கியத்தின் பகுத்துக் குறித்தலுக்கு முக்கியமாகும். பின்வரும் அட்டவணை சொல் வகைப்பாடுகளையும் அவற்றுடன் தொடர்புடைய அடையாளங்களையும் அவற்றின் பண்புக்கூறுகளுடன் காட்டும்.

சொல் வகைப்பாட்டுடன் தொடர்புடைய அடையாளப்படுத்திகள்

வ.எண்	சொல் வகைப்பாடு	அடையாளம்/உருப்படுத்தம்	எடுத்துக்காட்டு
1.	முற்றுவினை	V	பார்த்தான்
2.	பெயர்	N	பையன்
3.	மாற்றுப்பெயர்	P	அவர்
4.	பெயரடை	J	நல்ல
5.	வினையடை	R	வேகமாக
6.	பின்னுருபு	I	முன்னால்
7.	சுட்டடை	D	அந்த

8.	அளவடை	G	பல
9.	பெயர் மாறுபடுத்தி	M	புதுமை
10.	கிழமைப்பெயர்	GN	பொருளின்
11.	முற்றுப்பெறாத வினை	NV	பெற்று
12.	பெயரெச்ச வினை	RV	படித்த
13.	நிரப்பிய வினை	CV	என்று
14.	செயல்பாட்டுச் சொல்	F	வரை

5.5.2.3.2 இலக்கணப் பண்புகளும் அவற்றுடன் தொடர்புடைய அடையாளப்படுத்திகளும்

வ.எண்	பண்புகள்	சொல்வரை/உருப்படுத்தம்
பால்		
1.	ஆண்பால்	+M, -F
2.	பெண்பால்	-M, +F
3.	பலர்பால்	+M, +F
4.	ஒன்றன் பால்	-M, -F
எண்		
5.	ஒருமை	+S
6.	பன்மை	-S
இடம்		
7.	தன்மை	+1, -2
8.	முன்னிலை	-1, +2
9.	படர்க்கை	+1, +2

காலம்		
10.	நிகழ்காலம்	+P, -A
11.	இறந்தகாலம்	-P, +A
12.	எதிர்காலம்	+P, +A
வேற்றுமை		
13.	செயப்படுபொருள் வேற்றுமை	+ACC
14.	கொடை வேற்றுமை	+DAT
15.	இட வேற்றுமை	+LOC
16.	கருவி வேற்றுமை	+INS
17.	உடனிகழ்ச்சி வேற்றுமை	+SOC
18.	நீங்கல் வேற்றுமை	+ABL
19.	பயன் வேற்றுமை	+BEN
20.	கிழமை வேற்றுமை	+GEN

7.4.3.3 உருபனியல் பகுப்பாய்வு

உருபனியல் பகுப்பாய்வு சொற்களைப் பகுக்கும் செயல்பாட்டையும் அவற்றின் ஒட்டு உருபன்பளையும் வேர்களைக் கண்டுபிடிக்கும் செயல்பாட்டையும் உள்ளடக்கும். இவ்வுருபனியல் பகுப்பாய்வினால் ஒரு சொல்லின் சொல் வகைப்பாட்டையும் அதன் இலக்கணப் பண்புக்கூறுகளையும் அடையாளம் காண இயலும்.

7.4.3.4 சொல் வடிவமும் அதன் அடையாளங்களும்

சொல்வடிவம்	உருபன் விடுவரல்	சொல்லும் அதன் அடையாளப்படுத்தியும்

கேட்டான்	கேள்<verb>+த்த்<past tense>+ஆன்<3rd person, singular,masculine>	கேட்டான் <V:+p,-A, +1.+2 +3, +S, +M, -F>
மரத்தை	மரம் <noun> + ஐ <accusative case>	மரத்தை <N:+ACC, +S, -M, -F>

7.4.3.5 சொல்வகை அடையாளப்படுத்தல்

சொல்வகை அடையாளப்படுத்தல் வாக்கியங்களின் உறுப்புகளுக்குச் சரியான அடையாளத்தைத் தரும் செயல்பாடாகும். தமிழ் போன்ற மொழிகளுக்கு உருபனியல் பகுப்பாய்வி 70% துல்லியத்துடன் சொல்வகைப்பாட்டை கண்டுபிடிக்க இயலும். இருப்பினும் பல நிகழ்வுகளில் உருபனியல் பகுப்பாய்வி ஒரு சொல்லுக்கு ஒரு அடையாளத்தை மயக்கமின்றி தர இயலாது. இம்மாதிரியான மயக்கங்களை நீக்கக் குமாரசண்முகம் உருபனியல் பகுப்பாய்விற்குப் பின்னர் சொல்வகை அடையாளப்படுத்தியைப் பயன்படுத்துகிறார். சொல்வகை அடையாளப்படுத்தி உருபனியல் ஆய்வியின் விடுவரல் (output) ஏற்று இயலுகின்ற அளவு சொல்மட்ட விதிகளைப் பயன்படுத்தி மயக்கத்தை நீக்குகிறது. எஞ்சுவது பகுத்துக்குறிப்பானால் கவனித்துக் கொள்ளப்படும்.

7.4.3.6 இயல்பாக்கம் (Normalization)

தமிழ் ஒத்தறி அடிப்படையில் சொல்நிரல் சுதந்திரமுள்ள மொழியாகையால் இறுதி வினைக்கு முன்னர் வரும் பெ.தொ பங்கெடுப்பாளர்கள் இறுதி வினைக்கு முன் எந்த வரிசையிலும் வந்து ஒரே பொருளைத் தரும்.

எ.கா

ஆசிரியர் குழந்தைக்குப் பரிசைக் கொடுத்தார்.

ஆசிரியர் பரிசைக் குழந்தைக்குக் கொடுத்தார்.

பரிசை ஆசிரியர் குழந்தைக்குக் கொடுத்தார்.

பரிசைக் குழந்தைக்கு ஆசிரியர் கொடுத்தார்.

குழந்தைக்குப் பரிசை ஆசிரியர் கொடுத்தார்.

குழந்தைக்கு ஆசிரியர் பரிசைக் கொடுத்தார்.

குமாரசண்முகம் இயல்பாக்கத்திற்காக வாக்கியங்களில் தொடர்கள் ஒரு அனுமதிக்கப்பட்ட எல்லைக்குள் ஏற்கெனவே தீர்மானிக்கப்பட்ட வரிசைகளில் வருவதாக எடுத்துக் கொள்கிறார். இயல்பாக்கி வாக்கியத்தை வரிசையாக ஆய்ந்து தொடர் எல்லைகளை அடையாளம் காண்கின்றது. அதைத் தொடர்ந்து அது எச்சத் தொடர்களையும், பெயர்த்தொடர்களையும் இறுதி வினைக்கு முன் ஒரு குறிப்பிட்ட நிலையில் வரிசைப்படுத்துகிறது. பெயர்த்தொடர்களின் மற்றும் எச்சத் தொடர்களின் எல்லைகள் கண்டுபிடிக்கப்பட்ட பின் இயல்பாக்கி இவற்றை வலமிருந்து இடமாக பின்வரும் முன்நிர்ணயிக்கப்பட்ட நிரலில் மறுவரிசைப்படுத்தும்.

0. வினை,முற்றுப்பெறா வினை,பெயரெச்ச வினை,நிரப்பு வினை

1. பெ.தொ+செயப்படுபொருள் வேற்றுமை
2. பெ.தொ+கொடை வேற்றுமை
3. பின்னுருபு+இட வேற்றுமை
4. பின்னுருபு+கருவி வேற்றுமை
5. பின்னுருபு+உடனிகழ்ச்சி வேற்றுமை
6. பின்னுருபு+நீங்கல் வேற்றுமை
7. பின்னுருபு+பயன் வேற்றுமை
8. வினையடை+கிழமை வேற்றுமை

9. எழுவாய்

எடுத்துக்காட்டாக முன்னர் பட்டியலிடப்பட்ட வாக்கியங்களில் எது தரப்பட்டாலும் இயல்பாக்கி பின்வரும் வெளிப்பீட்டைத் தரும்.

ஆசிரியர் குழந்தைகளுக்குப் பரிசைக் கொடுத்தார்.

7.4.3.7 இலக்கணச்சட்டகம்

பகுத்துக் குறித்தலுக்குக் குமாரசண்முகம் தொடரமைப்பு இலக்கணத்தைப் பின்பற்றுகிறார். சொல் வகைப்பாடுகளில் இருந்து கிளைகளை உருவாக்க X-பார் கோட்பாட்டை ஒக்கும் அணுகுமுறையைப் பின்பற்றுகிறார்.

தொடரமைப்பு விதிகள் ஒரு குழும விதிகளைக் கொண்டது. சண்முகசுந்தரம் பின்பற்றுகின்ற விதிகள் கீழே பட்டியலிடப்பட்டுள்ளன.

Sl.No	Rules	Examples

	Quantifier Phrase Rules	Q->Quantifier
1	Q '=> Q	Q(Q(சில)) Q(Q(few))
2	QP => Q	QP (Q(சில)) QP (Q(few))
	Adjectival Pharase (ADJP) Rules	j->Adjective
3	J => J	J'(J(நல்ல)) J'(J(good))
4	ADJP => J	ADJP(நல்ல)) ADJP(J'(good))
	Modifier Phrase (MP) Rules	M->Noun modifier
5	M => M	M'(M'(கலை)) M'(M'(art))
6	MP => M	MP'(MP'(கலை)) MP (MP'(art))
	Genitive Phrase (GP) Rules	GN->Genitive noun D->Determiner RC->Relative Clause

7	GN => *GN	GN'(GN'(மாணவனுடைய)) GN'(GN'(student's))
8	N => MP*GN	GN'(MP(கல்லூரி) GN'(மாணவனுடைய)) GN'(MP'(college)GN'(student's))
9	GN => ADJP*GN	GN'(ADJP (நல்ல) GN'(மாணவனுடைய)) GN'(ADJP(good) GN'(student's))
10	GN'=> RC*GN	GN'(RC(<trace>படத்தைப் பார்த்த) GN'(பையனுடைய)) RC(<trace>movie saw+RP)GN'(boy's))
11	GN'=> GP*GN	GN'GP(மாணவனுடைய)GN'(புத்தகத்தின்)) GN'(GP(student's) GN'(book's))
12	GP=>*GN	GN'(GP' மாணவனுடைய)) GP'(GP'(student's))
13	GP=>D*GN'	GP(D(அந்த)GN'(மாணவனுடைய)) GP(D(that)GN'(student's))
	Noun Phrase (NP) Rules	P-> Pronoun N-> Noun F-> Functional word
14	N' => *N	N'(N(கல்லூரி)) N'(N(college))
15	N' => P *N	N'(P(அவன்) N'(P(he))
16	N' => MP *N'	N'(MP(fiy)N'(கல்லூரி)) N'(MP(art)N'(collegep))
17	N' => QP *N'	N'(QP(சில)N'(கல்லூரி)) N'(QP(few)N'(college))

18	N' => ADJP*N'	N'(ADJP(நல்ல)N'(கல்லூரி)) N'(ADJP(good)N'(college))
19	N' => RC*N'	N'(RC(<tarace> படத்தைப் பார்த்த)N'(பையன்)) N'(RC(<tarace>movie saw+RP)N'(boy))
20	N' => GP*N'	N'(GP(மாணவனுடைய)N'(புத்தகம்)) N'(GP(student's)N'(book))
21	N' => N'F	N'(N'(வீடு)F(வரை)) N'(N'(house)F(till))
22	NP => *N'	N'(N'(கல்லூரி)) N'(N'(college))
23	NP => D*N'	NP(D(அந்த)N'(கல்லூரி)) NP(D(that)N'(college))
	Postposition Phrase (PP) Rules	I -> Postposition N' -> N'wutg itger tgan Accusative and Dative cases
24	PP => *N'	PP(N'(கல்லூரியில்)) PP(N'(college in))
25	PP => D*N'	PP(D(அந்த)N'(கல்லூரியில்)) PP(D(that)N'(college in))
26	P => *N'I	PP(N'(வீட்டிற்கு)I(முன்)) (N'(house in)I(front of))
	Adverbial Pharase (ADVP) Rules	R -> Adverb
27	R' => R	R'(R(மெதுவாக))

		R’(R(slowly))
28	R’=> QPR’	R’(QP(மிக)R’(மெதுவாக)) R’(QP(very)R’(slowly))
29	ADVP => R’	ADVP(R’(மெதுவாக)) ADVP(R’(slowly))
	Verrb Phrase (VP) Rules	V -> Verb
30	V’ => *V	V’(V(பார்த்தான்)) V’(V(saw))
31	V’ => NP *V	V’PP(காடுகளில்)V(காணப்படுகிறது)) V’PP(in forest)V(seen))
32	V’ => PP*V	V’(NP(படத்தைப்)V(பார்த்தான்)) V’(NP(movie)V(saw))
33	V’ => PP PP *V	V’(ஜோடியுடன்PP(காடுகளில் V(காணப்படுகிறது)) V’(with pair)PP(in forest)V(seen))
34	V’ => NP NP *V	V’(NP(அவனுக்கு)NP(புத்தகத்தைக்)V(கொடுத்தேன்)) V’(NP(him)NP(book)V(gave))
35	V’ => PP NP *V	V’(PP(திரையரங்கில்)NP(படத்தைப்)V(பார்த்தான்)) V’(PP(in theatre)NP(movie)V(saw))
36	V’ => PP PP NP *V	V’(PP(திரையரங்கில்)PP(நண்பருடன்)NP((படத்தைப்)V(பார்த்தான்)) V’(PP(in theatre)PP(with friend)NP(movie)V(saw))
37	V’ => PP NP NP *V	V’(PP(வீட்டில்)NP(அவனுக்கு)NP(புத்தகத்தைக்)V(கொடுத்தேன்)) V’(PP(at home)NP(him)NP(book)V(gave))
38	V’ => NC *V’	V’(NC(படத்தைப் பார்த்து)V’(கருத்தைச் சொன்னான்))

		V'(NC(movie saw+VBP)V'(comments said))
39	V' => ADVP *V'	V'(ADVP(கவனமாகப்V'(பார்த்தான்)) V'(ADVP(carfully)V'(saw))
40	V' => *V'	VP(V'(பார்த்தான்)) VP(V'(saw))
	Nonfinite Verb Phrase (NVP) Rules	NV -> Nonfinite Verb NC->Non-finite clause VBP-> Verbal Participial
41	NV' => *NV	NV'(NV(பார்த்து)) NV'(NV(saw))
42	NV' => PP *NV	NV'(PP(காடுகளில்NV(காணப்பட்டு)) NV'(PP(in forest)NV(seen+VBP))
43	NV' => NP *NV	NV'(NP(படத்தைப்)NV(பார்த்து)) NV'(NP(movie)NV(saw+VBP))
44	NV' => PP PP *NV	NV'(PP(ஜோடியுடன்)PP(காடுகளில்)NV(காணப்பட்டு)) NV'(PP(with pair)PP(in forest)NV(seen+VBP))
45	NV' => NP NP *NV	NV'(NP அவனுக்கு)NP(புத்தகத்தைக்)NV(கொடுத்து)) NV'(NP(him)NP(book)NV(gave+VBP))
46	NV' => PP NP *NV	NV'(PP(திரையரங்கில்)PP(படத்தைப்)NV(பார்த்து)) NV'(PP(in theatre)PP(movie)NV(saw+VBP))
47	NV' => PP PP NP *NV	NV'(PP(திரையரங்கில்)PP(நண்பனுடன்PP(படத்தைப்)NV(பார்த்து)) NV'(PP(in theatre)P P(withfriend)PP(movie)NV(saw+VBP))
48	NV' => PP NP NP *NV	NV'(PP(வீட்டில்NP(அவனுக்கு)NP(புத்தகத்தைக்)NV(கொடுத்து)) NV'(PP(at home)NP(him)NP(book)NV(gave+BVP))
49	NV' => NC *NV'	NV'(NC(படத்தைப் பார்த்து)NV'(கருத்தைச் சொல்லி)) NV'(NC(movie saw+VBP)NV'(comments said+VBP))

50	NV' => ADVP *NV'	NV'(ADVP(கவனமாகப்)NV'(பார்த்து)) NV'(ADVP(carefully)NV'(saw+VBP))
51	NV' => *NV'	NVP(NV'(பார்த்து)) NVP(NV'(saw_VBP))
	Nonfinite Clause (NC) Rules	CV->Complementizing Verb
52	NC => *NP *NVP'	NC'(NP(அவன்)NVP(படத்தைப் பார்த்து)) NC'(NP(he)NVP(movie saw+VBP))
53	NVP => *NV'	NC(S(அவன் படத்தைப் பார்த்தான்)CVஎன்று)) NC(S(he movie saw)CV(that))
	Relativizing Verb (RVP) Rules	RV -> Relativizing Berb RP -> Relative Participle
54	RV' => *RV	RV'(RV(பார்த்த)) RV'(RV(saw+RP))
55	RV' => PP *RV	RV'(PP(காடுகளில்)RV(காணப்பட்ட)) RV'(PP(in forest)RV(seen+RP))
56	RV' => NP *RV	RV'(NP(படத்தைப்RV(பார்த்த)) RV'(NP(movie)RV(saw+RP))
57	RV' => PP PP *RV	RV'(PP(ஜோடியுடன்)PP(காடுகளில்)RV(காணப்பட்ட)) RV'(PP(with pair)PP(in forest)RV(seen+RP))
58	RV' => NP NP *RV	RV'(NP(அவனுக்கு)NP(புத்தகத்தைக்)RV(கொடுத்த)) RV'(NP(him)NP(book)RV(gave+RP))
59	RV' => P NPP *RV	RV'(PP(திரையரங்கில்)PP(படத்தைப்)RV(பார்த்த)) RV'(PP(in theatre)PP(movie)RV(saw+RP))

60	RV' => PP PP NP *RV	RV'(PP(திரையரங்கில்)PP(நண்பனுடன்)PP(படத்தைப்)RV(பார்த்த)) RV'(PP(in theatre)P P(withfriend)PP(movie)RV(saw+RP))
61	RV' => PP NP NP *RV	RV'(PP(வீட்டில்)NP(அவனுக்கு)NP(புத்தகத்தைக்)RV(கொடுத்த)) RV'(PP(at home)NP(him)NP(book)RV(gave+RP))
62	RV' => NC *RV'	RV'(NC(படத்தைப் பார்த்து)RV'(கருத்தைச் சொன்ன)) RV'(NC(movie saw+VBP)RV'(comments said+RP))
63	RV' => ADVP *RV'	RV'(ADVP(கவனமாகப்RV'(பார்த்த)) RV'(ADVP(carefully)RV'(saw+RP))
64	RV' => *RV'	RVP(RV'(பார்த்த)) RVP(RV'(saw+RP))
	Relative Clause (RC) Rules	
65	RC => *NP *VP	RC(NP(<trace>)RVP(படத்தைப் பார்த்த)) RC(NP(<trace>)RVP(movie saw+P))
	Sentence (S) Rules	
66	S => *NP *VP	S(NP(mtd)VP(அவன் படத்தைப் பார்த்தான்)) S(NP(he)VP(movie saw))

ஒரு விதியின் தலைப்பு உறுப்பு 'என்ற குறியீட்டால் முன் தொடரப்படும் வாக்கியம் பகுத்துக் குறிக்கப்படும்போது தலைமை உறுப்பின் இலக்கண பண்புக்கூறுகள் இடப்பக்கமுள்ள இறுதியுறாத ஒன்றிற்குக் கடத்தப்படுகிறது. M என்ற சொல் வகைப்பாடு கூட்டுப்பெயர்களைக் கையாளவேண்டி பயன்படுத்தப்படுகிற தனித்தன்மையான வகைப்பாடாகும். 'M' என்பது மாறுபடுத்தும் தொடரியல் (Modifier Phrase (MP)) தன்மை உறுப்பாகும். கிழமை வேற்றுமைக் குறியீடு உள்ள ஒரு பெயர் மட்டும்தான் கிழமைத்தொடரின் (Genitive Phrase/GP) தலைமையாக வரும். பின்னுருபுத் தொடர்கள் பெயரைத் தொடர்ந்து முன்,பின் என்ற பின்னுருபுகளால்

மெய்ப்படுத்தம் செய்யப்படும். முற்றுப்பெறாத எச்சத்தொடரில் (Non-finite Clause/NC) தலைமை ஒரு முற்றுப்பெறாத வினையாகும். தமிழில் நான்கு வகையான முற்றுப்பெறா வினைகளிலிருந்து நான்கு வகையான முற்றுப்பெறா எச்சத்தொடர்கள் கிடைக்கின்றன. எச்சத்தொடர் (Relative clause (RC)) பெயர்களுக்கு முன்னர் வந்து அவற்றை அடை செய்யும்.

7.6 இந்தியமொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர்க்கூறு பகுப்பாய்வு

இந்தியமொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர்க்கூறு பகுத்தல் ஒருவாக்கியத்தின் தனிநிலை பெயர்தொடர்கள், வினைக் குழுமங்கள், பெயரடைத் தொடர்கள், வினையடைத் தொடர்கள் இவற்றை அடையாளங்காணுதலைக் குறிப்பிடும். இது தொடர் உறுப்புகளின் எல்லையையும் புலக்குறிப்பையும் கண்டுபிடிப்பதை உட்படுத்தும்.

இடுவரல்-விடுவரல் தனிக்குறிப்பீடுகள்

Input-Output Specification

Input: TKN_, CAT_

Output: ADDR_ (ADDR என்பதன் மதிப்பு மாறுகையில் மீளமைப்புச் செயல்பாடு நடக்கிறது என்பது இதன் அர்த்தமாகும்.)

மேலே தரப்பட்டுள்ள இடுவரல் தனிக்குறிப்பீடுகள் SSF-இல் வருணனை செய்வதற்கு TKN_ மற்றும் CAT_ என்ற தனிப்பண்புகள் அதற்குத் தேவை அல்லது அது பன்படுத்துகின்றது என்று அர்த்தம்தரும். வேறுவிதமாகக் கூறினால் இடுவரல் வாக்கியத்தில் சொற்கள் அல்லது டொக்கன்கள் சொல்வகைப்பாட்டு வகையுடன் தரப்பட்டுள்ளன.

மேலே தரப்பட்டுள்ள விடுவரல் தனிக்குறிப்பீடுகள் 'ADDR_' இந்தத் தொகுதியால் வரையறவிளக்கம் பெறும் என்று அர்த்தம் பெறும்.

எடுத்துக்காட்டு:

வருணனை: தனிநிலைப் பெயர்த்தொடர்கள் தொடர் உறுப்புகளாகக் (NPs) குழுமப்படும். வினைத் தொடர்ச்சிகள் வினைத் தொடர் உறுப்புகளாகக் (VG) குழுமப்படும்.

7.6.1 இ-இ இயந்திர மொழிபெயர்ப்புத் திட்டத்திற்குத் தேர்ந்தெடுக்கப்பட்டுள்ள தொடர்கூறுகள்

இதில் தொடர் உறுப்புகளின் வரையறை விளக்கமும் தொடர் உறுப்பு வகைகளும் சில தனித்துவம் வாய்ந்த நேர்வுகளும் பேசப்படும். இந்திய மொழிகளுக்கு இடையிலான

மொழிபெயர்ப்புத் திட்டத்தில் 11 தொடர் உறுப்புகள் தெரிந்தெடுக்கப்படு அவை தொடர் உறுப்பு அடையாளப் படுத்தலுக்குப் பயன் படுத்தபடுகின்றன.

முக்கிய தொடர்கூறு குழுமம் அவற்றின் உருப்படுத்தமும்

	முக்கிய தொடர்கூறு குழுமம்	உருப்படுத்தம்
1	பெயர்த் தொடர்கூறு	NP
2	வினைமுற்றுத் தொடர்கூறு	VGF
3	வினைமுற்றாத் தொடர்கூறு	VGNF
4	வினைப்பெயர்த் தொடர்கூறு	VGNN
5	வினை எச்சத்தொடர்கூறு	VGINF
6	பெயரைடைத் தொடர்கூறு	JJP
7	வினையடைத் தொடர்கூறு	RBP
8	இணைப்புத் தொடர்கூறு	CCP
9	வியப்புத் தொடர்கூறு	INTJ
10	எதிர்மறைத் தொடர்கூறு	NEGP
11	வெற்றிடத் தொடர்கூறு	BLK

7.6.1.1 பெயர்த் தொடர்கூறு (NP)

பெயர்த்தொடர் உறுப்புகள் (NP chunks) NP என அடையாளப் படுத்தப்படும்; இது மறுதரவாக வராத பெயர்த்தொடர்களையும் பின்னுருபுத் தொடர்களையும் உட்படுத்தும். பெயர்த்தொடர் உறுப்பின் தலைமை ஒரு பெயராகும். சிறப்புசெய்யும் சொற்கள் பெயர்த்தொடர் உறுப்பின் இடது பக்க எல்லையை நிறுவும்; வேற்றுமையோ தலைமைப் பெயரோ வலது பக்க எல்லையை நிறுவும். பெயரை அடைசெய்யும் வருணனைப் பெயரடை(கள்) பெயர்த்தொடர் உறுப்பின் பகுதியாக அமையும். தலைமைப் பெயருடன் நங்கூரமிட்டுள்ள குறைச் சொற்களும் பெயர்த்தொடர் உறுப்புடன் குழுமப்படும். அது பெயருக்குப் பின்னரோ வேற்றுமைக்குப் பின்னரோ வந்தால், அது தொடர் உறுப்பின் வலது பக்க எல்லையை நிர்ணயிக்கும். பெயர்த்தொடர் உறுப்புகளின் சில எடுத்துக்காட்டுகள் கீழே தரப்பட்டுள்ளன.

((குழந்தைகள்_NN))_NP, ((சில_QF குழந்தைகள்))_NP, ((சில_QF நல்ல_JJ குழந்தைகள்_NN))_NP, ((மேசை மேல்_PSP))_NP, ((ஒரு_QC கறுத்த_JJ குதிரை_NN))_NP, ((இந்த_DEM புதிய_JJ புத்தகம்_NN))_NP, ((இந்த_DEM புதிய_JJ புத்தகத்தின்_NN மேல்_PSP))_NP

7.6.1.2 வினைத் தொடர்கூறு (VG)

வினைத் தொடர் உறுப்புகள் பலவகைப்படும். ஒரு வினைக் குழுமம் முக்கிய வினையையும் இருப்பின் துணை வினைகளையும் உள்ளடக்கும். பின் வருவது வினைத் தொடர் உறுப்புகளின் எடுத்துக்காட்டுகளாகும்.

((போனான்)), ((போய்விட்டான்)), (போய்க்கொண்டிருக்கிறான்))

7.6.1.3 பெயரைடைத் தொடர்கூறு (JJP)

பெயரடைத் தொடர் உறுப்பு JJP என அடையாளப்படுத்தப்படும். இத்தொடர் பயனைலைப் பெயர்தொடர்களையும் உள்ளடக்கி எல்லாப் பெயர்த்தொடர் உறுப்புகளையும் கொண்டிருக்கும். இருப்பினும் பெயருக்கு முன்னர் வரும் பெயரடைகள் பெயர்த் தொடர் உறுப்புடன் சேர்த்துக் குழுமப்படும்.

7.6.1.4 வினையடைத் தொடர்கூறு (RBP)

இத்தொடர் உறுப்பின் பெயரும் சொல்வகை அடையாளப் படுத்தலின் போது பயன்படுத்தப்படுகிற சொல்வகை அடையாளத்தின் அடிப்படையில் அமைகின்றது. எல்லா வினையடைத்தொடர் உறுப்பும் RBP என அடையாளப்படுத்தப்படும். இந்தத் தொடர் உறுப்பு எல்லாச் சுத்தமான வினைத்தொடர்களையும் உட்படுத்தும்.

7.6.1.5. இணைப்புத் தொடர்கூறு (CCP)

இணைப்புத் தொடர் உறுப்புகள் யாவும் CCP என அடையாளப்படுத்தப்படும். இத்தகவல் பெரிய வாக்கிய அமைபுகளை உருவாக்கத் தேவையாகும்.

7.6.1.6 வியப்புத் தொடர்கூறு (INTJ)

எல்லா வியப்புத் தொடர் உறுப்புகளும் INTJ என அடையாளப்படுத்தப்படும்.

7.6.1.7 எதிர்மறைத் தொடர்கூறு (NEGP)

வினையைச் சுற்றி இருக்கும் எதிர்மறைக் குறைச்சொல் வினைக் குழுவுடன் குழுமப்படவேண்டும். அது வினையிலிருந்தோ பெயரிலிருந்தோ விலகி இருந்தால் அது தனியான தொடர் உறுப்பாக குழுமப்படவேண்டும்.

7.5.2. தொடர்கூறு பகுத்தலின் எடுதுக்காட்டுகள்

AU-KBC-நிறுவனதில் நடைபெற்றுவரும் இந்திய மொழியிலிருந்து இந்திய மொழிபெயப்பு (தமிழ்-இந்தி) ஒழுங்குமுறையின் இரு கூறுகள் இவ்வாய்விற்கு எடுத்துக்கொள்ளப்பட்டுளன என்பதன் அடிப்படையில் இவ்விரு செயல்பாடுகளையும் வெளிப்படுத்தும் பொருட்டுத் தமிழ் தரவுத் தொகுதியில் எவ்வாறு சொல்வகை அடையாளப்படுத்தலும் தொடர் உறுப்பு பிரித்தலும் நடைபெறுகின்றன என்பதைக் கூறும் முகமாகக் கீழே எடுத்துக்காட்டு தரப்பட்டுள்ளது.

<NP> தலம்/NN :/SYM </NP> <NP> வைணவ/JJ தலம்/NN </NP>

<NP> அமைவிடம்/NN :/SYM </NP> <NP> திருச்சியிலிருந்து/NN </NP> <NP> 3/QC மைல்கள்/NN தூரம்/NN </NP>

<NP> திருக்கோயிலின்/NN பெயர்/NN :/SYM </NP> <NP> ஸ்ரீ/NNC அரங்கநாதசுவாமி/NNP திருக்கோயில்/NN </NP>

<NP> இறைவன்/NN /SYMஇறைவி/NN திருநாமங்கள்/NN :/SYM </NP> <NP> ஸ்ரீ/NNC அரங்கநாத/NNPC சுவாமி/NNP </NP> <NP> ஸ்ரீ/NNC ரங்கநாச்சியார்/NNP </NP>

<NP> தீர்த்தங்கள்/NN :/SYM </NP> <NP> சந்திரபுஷ்கரணி/NN </NP> <CCP> மற்றும்/CC </CCP> <NP> சூரிய/JJ புஷ்கரணி/NN </NP>

<NP> முக்கிய/JJ விழாக்கள்/NN :/SYM </NP> <NP> வைகுண்ட/NNC ஏகாதசி/NN பிரம்மோற்சவம்/NN </NP>

<NP> சிறப்பு/NN அம்சம்/NN :/SYM </NP> <NP> 108/QC வைணவத்/JJ திருத்தலங்களில்/NN </NP> <JJP> முதன்மையான/JJ </JJP> <VGF>திருக்கோயிலாகும்/VM ./SYM </VGF>

<NP> 12/QC ஆழ்வார்களில்/NN </NP> <NP> 11/QC ஆழ்வார்கள்/NN </NP> <NP> இதன்/PRPபுகழை/NN </NP> <NP> பெருமையை/NN </NP> <NP> கீர்த்தியைப்/NN </NP> <VGF>பாடியுள்ளார்கள்/VM ./SYM </VGF>

<NP> இந்த/DEMதலத்தில்தான்/NN </NP> <NP> ஆண்டாள்/NNP </NP> <NP> திருப்பாணாழ்வார்/NNP </NP> <NP> பீபிநாச்சியார்/NNP </NP> <CCP> மற்றும்/CC </CCP> <NP> பலர்/NN </NP> <NP> அரங்கனின்/NNP திருவடியை/NN </NP> <VGF>அடைந்தனர்/VM ./SYM </VGF>

<RBP> சுயம்புவாக/RB </RBP> <NP> உருவான/JJ ஒரே/QC திருக்கோயில்/NN ./SYM </NP>

<NP> போக்குவரத்து/NN மையம்/NN :/SYM </NP> <NP> திருச்சி/NNP ./SYM </NP>

<NP> அரங்கன்/NNP </NP> <VGNF> குடிகொண்டுள்ள/VM </VGNF> <NP> திருவரங்கம்/NNP </NP<NP> திருச்சி/NNP /SYM</NP> <NP> சென்னை/NNP கார்டு/NN ரெயில்/NN பாதையில்/NN </NP> <BLK> உள்ள/PSP </BLK> <NP> ரேVயில்/NN </NP> <VGF>நிலையமாகும்/VM ./SYM </VGF>

<NP> சென்னையிலிருந்து/NNP </NP> <NP> 207/QC மைல்கள்/NN தூரத்திலும்/NN </NP> <NP> திருச்சிராப்பள்ளி/NNP சந்திப்பிலிருந்து/NN </NP> <NP> 3/QC மைல்கள்/NN தொலைவிலும்/NN </NP> <VGF>உள்ளது/VM ./SYM </VGF>

<NP> திருவரங்கம்/NNP நகரம்/NN </NP> <NP> திருச்சிராப்பள்ளி/NNP மாநகரின்/NN புறநகரில்/NN </NP> <BLK> உள்ள/PSP </BLK> <JJP> சிறிய/JJ </JJP> <VGF>நகரமாகும்/VM ./SYM </VGF>

<NP> ஒருபுறம்/NN காவிரியும்/NNP </NP> <NP> மற்றொருபுறம்/NN கொள்ளிடமும்/NNP </NP> <VGNF>பாய்ந்தோடும்/VM </VGNF><NP> இந்த/DEMதிருத்தலத்துக்கு/NN </NP> <RBP> ஆண்டுமுழுவதும்/RB </RBP> <NP> ஆயிரக்கணக்கான/JJ இந்துக்கள்/NN </NP> <VGNF>வந்து/VM </VGNF><NP> அரங்கனை/NNP </NP> <VGF>வழிபட்டு/VM செல்கின்றனர்/VAUX ./SYM </VGF

<RBP> குறிப்பாக/RB </RBP> <NP> வைஷ்ணவர்களுக்கு/NN </NP> <NP> இதுPSP</NP> <JJP> முக்கியமான/JJ </JJP> <VGF>இடமாகும்/VM ./SYM </VGF>

இயல் 8

முடிவுரை

இயந்திர மொழிபெயர்ப்பு காலப்போக்கில் பலவித மாற்றங்களையும் செயல்பாடுகளையும் தாண்டி வந்துள்ளது. இன்றளவும் ஒரு முழுமையான தானியங்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கப்படவில்லை. மேற்கண்ட இயல்கள் இதை நமக்கு அறிவுறுத்தும். தமிழை மையமாகக்கொண்டு எழுதப்பட்ட இயந்திர மொழிபெயர்ப்பு குறித்த இந்தூல் ஒரு முழுமையான படைப்பு என்று கூறவியலாது.

இயந்திர மொழிபெயர்ப்பு என்பது ஒரு மூல மொழியில் இருந்து ஒரு உரையை இலக்கு மொழியில் மொழிபெயர்க்கும் செயல்பாடாகும். இயந்திர மொழிபெயர்ப்பில் பல சவாலான அம்சங்கள் உள்ளன: (1) பல்வேறு வகையான மொழிகள், எழுத்துக்கள் மற்றும் இலக்கணங்கள்; (2) ஒரு வரிசையை (எடுத்துக்காட்டாக ஒரு வாக்கியம்) ஒரு வரிசைக்கு மொழிபெயர்க்கும் செயல்பாடு எண்களுடன் மட்டுமே செயல்பாடு செய்வதை விட கணினிக்கு கடினம்; (3) சரியான பதில் எதுவும் இல்லாமை (எ.கா.: பாலினம் சார்ந்த பதிலீட்டுப்பெயர்கள் இல்லாமல் ஒரு மொழியிலிருந்து மொழிபெயர்ப்பது). அவரும் அவளும் ஒரே மாதிரியாக இருக்க இயலும்.

இயந்திர மொழிபெயர்ப்பு ஒப்பீட்டளவில் பழைய பணியாகும். 1970களில் இருந்து, தானியங்கி மொழிபெயர்ப்பை அடைவதற்கான திட்டங்கள் இருந்தன. பல ஆண்டுகளாக, மூன்று முக்கிய அணுகுமுறைகள் தோன்றின:

• விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Rule-based Machine Translation (/RBMT/ஆர் எம்டி): 1970களிலிருந்து 1970கள் வரை

• புள்ளிவிவர இயந்திர மொழிபெயர்ப்பு (Statistical Machine Translation (SMT/எஸ்எம்டி): 1990 களிலிருந்து 2010கள் வரை

• நரம்பியல் இயந்திர மொழிபெயர்ப்பு (Neural Machine Translation (NMT/ என்எம்டி): 2014-

துணைநூற்பட்டியல்

இராசேந்திரன். ச. & 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.

AAMT. 1993. MT Summit IV: proceedings. International cooperation for global communication. July 20-22, 1993.

AAMT. 1999. Machine translation summit VII '99: MT in the great translation era, September 13-17, 1999, Singapore.

AAMT. 2005. MT Summit IX: the Tenth Machine Translation Summit: proceedings, September 12-16, 2005, Phuket,Tailand.

AAMT. 1993. *MT Summit IV: proceedings. International cooperation for global communication*. July 20-22, 1993.

Aarts.J. and Meijs, W. (Eds.)1986. Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora. Amsterdam-Atlanta, G.A.: Rodopi.

Aditi Kalyani, Priti S. Sajja. 2015 A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies. International Journal of Computer Applications (0975 – 8887), Volume 121 – No.23, July 2015

Aijmer, K. and Alwood, 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 Jab Stvartvik. London: Longman.

Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, .1997. "ANUSAARAKA: Machine Translation in stages', Vivek, a quarterly in Artificial Intelligence, Vol. 10, No. 3, NCST Mumbai.

Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal. 2001. ANUSAARAKA:overcoming the language barrier in India", published in Anuvad: approaches to Translation.

Akshar Bharati and Amba Kulkarni. 2009. "Anusaaraka: An accessor-cum-Machine Translator". Department of Sanskrit Sudies, Univeristy of Hyderabad.

Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Lakshmi Bai. 2006. AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages, Technical Report, Language Technologies Research Centre IIIT, Hyderabad.

ALPAC. 1966. *Language and machines*: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee... Washington, DC: National Academy of Sciences.
Alshawi, H. (ed.) 1992. *The core language engine*. Cambridge, Mass.: The MIT Press.

ALPAC. 1966. Language and Machines: computers in translation and linguistics. A report by the Automatic Langauge Processing Adivisory Committee… Washington, DC: National Academy of Sciences.

Amano, S. 1982 Machine Translation Project at Toshiba Corporation. Technical note. R&D Center, Information Systems Laboratory, Toshiba Corporation, Kawasaki, Japan.

Ambati, V., and Rohini, U. (2007). A Hybrid Approach to EBMT for Indian Languages. ICON 2007.

Amruta Godase and Sharvari Govilkar. 2015. Machine Translation Development For Indian Languages And Its Approaches. International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2,April 2015.DOI : 10.5121/ijnlc.2015.4205 55.

AMTA. 1991. *Machine Translation Summit III*, July 1-4, 1991. Proceedings, program, contributed papers, panel statements.

AMTA. 1994. *Technology partnerships for crossing the language barrier*. Proceedings of the First Conference of the Association for Machine Translation in the Americas, 5-8 October 1994, Columbia, Maryland, USA

AMTA. 1996. *Expanding MT horizons*. Proceedings of the Second Conference of the Association for Machine
Translation in the Americas, 2-5 October 1996, Montreal, Quebec, Canada

AMTA. 1997. MT Summit VI – machine translation: past, present, future, ed. V. Teller & B. Sundheim 29 October- 1 November 1997, San Diego, California, USA.

AMTA. 1998. *Machine translation and the information soup. Third conference of the Association for Machine Translation in the Americas, AMTA '98, Langhorne, PA, USA, October 1998:* proceedings, ed. D.Farwell, L.Gerber & E.Hovy. Berlin: Springer.

AAMT. 1999. *Machine translation summit VII '99: MT in the great translation era*, September 13-17, 1999, Singapore.

AMTA. 2000. *Envisioning machine translation in the information future. 4th conference of the Association for Machine Translatio.*

AMTA. 2002. Machine translation: from research to real users. 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002, Tiburon, CA, USA, October 2002: proceedings, ed. S.D. Richardson. Berlin: Springer.

AMTA. 2003. *MT Summit IX. Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA,September 23-27, 2003.

AMTA. 2004. Machine translation: from real users to reseach: 6th conference of the Association for Machine Translation in the Americas, AMTA 2004, ed. R.E. Frederking and K.B. Taylor. Washington, DC, Spetember 2, 2004. Berlin: Springer Verlag.

AMTA. 2006. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas:*
Visions for the Future of Machine Translation, August 8-12, 2006, Cambridge, Mass., USA.

AMTA. 2006. Proceedings of 7th Conference of Association for Machine Translation in Americas: Vision for the Future of Machine Translation, August 8-12, 2006, Cambridge, Mass., USA>

Anand Kumar M, Dhanalakshmi V, Rekha R U,Soman K, P and Rajendran (2010). "A Novel data driven algorithm for Tamil morphological generator", IJCA,Vol.6,No.12,Pages:52-56.

Anand Kumar M, Dhanalakshmi V, Soman K.P , Rajendran S (2010) " A Sequence Labeling Approach to Morphological Analyzer for Tamil Language", IJCSE, Vol. 02, No. 06,, 1944-1951.

Anand Kumar M, Dhanalakshmi V, Soman K.P and Rajendran S (2010). "A Sequence Labelling Approach to Morphological Analyzer for Tamil Language". IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 1944-1951

Anand Kumar M, Dhanalakshmi V, Soman K P and Sharmiladevi V (2013). "Improving the Performance of English-Tamil Statistical Machine Translation System using Source-Side Pre-Processing", Proceedings of International Conference on Advances in Computer Science, AETACS © Elsevier, 2013

Anand Kumar M, Rajendran S. and Soman K. P. (2014) "AMRITA@ FIRE-2014: Morpheme Extraction for Tamil using Machine Learning (Working notes)", in International Workshop: "MET shared Task" Forum for Information Retrieval Evaluation (FIRE- 2014), Bengaluru , 2014.

Anand Kumar M, Rajendran S and Soman K P (2014) "Tamil Word Sense Disambiguation using Support Vector Machines with rich features." International Journal of Applied Engineering Research. ISSN 0973-4562, Volume X, Number X (2014) pp. xxx-xxx © Research India Publications http://www. ripublication.com/ijaer.htm

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, Rajendran S. and Soman K. P. (2015). "Cross-Lingual Preposition Disambiguation for Machine Translation." Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015). Available online at www.sciencedirect.com

Anand Kumar M, Dhanalakshmi V, Soman K.P. and Rajendran S. (2014). "Factored Statistical Machine Translation System for English to Tamil Language." Social Sciences & Humanities Journal homepage: ttp://www.pertanika.upm.edu.my/Pertanika J. Soc. Sci. & Hum. 22 (4): 1045 - 1061 (2014)

M Anand Kumar 2015. Tamil Linguistic Tools and English-Tamil Machine Translation System. United State: LAP Lambert Academic Publishing.

Anand Kumar M, S. Shriya, and K. P. Soman. (2015). AMRITA-CEN@FIRE 2015: Extracting entities for social media texts in Indian languages. CEUR Workshop Proceedings, 1587:85–88, 2015.

Anand Kumar M, Shivkaran Singh, Kavirajan B and Soman K P 2016. Shared Task on Detecting Paraphrases in Indian Languages (DPIL): An Overview. Anand Kumar M (http://orcid.org/0000-0003-0310-4510)

Anand Kumar M, Premjith B, Shivkaran Singh, Rajendran S and Soman K P. (2017). An Overview of the Shared Task on Machine Translation in Indian Languages (MTIL) - 2017 (Special issue on MTIL 2017) J. Intell. Syst. (), 1–16 DOI 10.1515/jisys-- © de Gruyter.

Aarts.J. and Meijs, W. (Eds.)1984. Corpus Linguistics: Recent Development in the use of Computer Corpora in English Language Research. Amsterdom-Atlanta. G.A.:Rodopi.

Ananthakrishna R. Kavitha M. Jayaprasad J. Hegde, Chandra Sekhar, Ritesh Shah, Sawani Bade and Sasikumar M. (2006). "MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine.

ARPA. 1994. *ARPA Workshop on machine translation, 17-18 March 1994, Sheraton Premier Hotel, Vienna, VA, USA.* [Washington, DC: ARPA]

Bar-Hillel, Y. 1960. "The present status of automatic translation of languages." Advances in Computers 1, 91-163.

Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal, (2003) "Machine Translation: The Shakti Approach", Pre-Conference Tutorial, ICON-2003.

Antony P. J., (2013) "Machine Translation Approaches and Survey for Indian Languages", International journal of Computational Linguistics and Chinese Language Processing Vol. 18,
No. 1, pp. 47-78.

Arulmozhi. P, Sobha L. 2006. A Hybrid POS Tagger for a Relatively Free Word Order Language. In proceedings of MSPIL-2006, Indian Institute of Technology, Bombay.

Atwell, E. (Eds.) 1993. Corpus Based Computational Linguistics. Amsterdam: Rodopi.

AU-KBC Post Tagset Tamil. AU-KBC Research Centre, MIT Campus of Anna University, Chrompet, Chennai-44.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bandyopadhyay, S. 2000. ANUBAAD – The Translator From English to Indian Languages. VIIth State Science and Technology Congree, Calcutta, India

Bernadini, S. 2000. Competence, Capacity, Corpora, Bologna: CLUEB.

Biber, D. Conrad, S. and Reppen, R. 1998. Corpus Linguistics-Investigating Language Structure and use. Cambridge University Press.

Booth, A.D. (ed.) 1967. Machine translation.Amsterdam: North-Holland.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R. and Roossin, P. 1988. "A statistical approach to French/English translation." In: TMI (1988).

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., … & Roossin, P. S. 1990. A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79–85.

Brown, P.F. Della Pietra, S.A., Della Pietra, V.J. and Mercer,R.L. 1993."The mathematics of statistical machine translation: parameter estimation." Computational Linguistics 19 (2) 263-311.

Carl,M. and Way, A.(eds.) 2003. Recent advances in example-based machine translation. Dordrechet: Kluwer. (Text, speech and Languge Technology, 21).

Chang,J.S. and Su, K.Y. 1997. "Corpus-based statistics-oreinted (CBSO) machine translation researches in Taiwan." In: AMTA (1997), 165-173.

CALTS in collaboration with, IIIT Hyderabad. English-Telugu T2T Machine Translation and Telugu-Tamil Machine translation System. Indo-German Workshop on Language technologies, AU-KBC Research Centre, Chennai, 2004 . www.au-kbc.org/dfki/igws/Machine_Translation.ppt.

CLAW. 1996. Proceedings of the First International Workshop on Controlled Language Applications, 26-27 March 1996, Centre for Computational Linguistics, Leuven. Leuven: Katholieke Universiteit.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

CDAC Mumbai, (2008) "MaTra: an English to Hindi Machine Translation System", a report by CDAC Mumbai formerly NCST.

Chomsky, A.N. 1968. Language and Mind. New York: Harcourt Brace.

Chomsky, A.N. 1972. Studies in Generative Grammar. The Hague: Mouton.

Clear, Jereny, 1992. Corpus sampling in New Directions in English Language Corpora (ed.) Gerhars Leitner. Mouton de Gruyter, New york.

CIIL, HINDI, CORPORA Tagset. Finalized on August 29, 2006.

Dasgupta, T., & Basu, A. (2008). An English to Indian Sign Language Machine Translation System, www.cse.iitd.ac.in/embedded/assistech/Proceedings/P17.pdf.

Darabari, Hemant. 1999. "Computer Assisted Translation System- An Indian Perspective", in proceedings of MT Summit VII, Thialand.

Dasgupta, T., Dandpat, S., & Basu, A. (2008). Prototype Machine Translation System From Text-To-Indian Sign Language. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, 19-26.

Dash, N.S. 2005. Corpus Linguistics and Language Technology (with reference to Indian Languages). New Delhi: Mitial Publications.

Dave, S., Parikh, J., & Bhattacharya, P. .2001. Interlingua-based English-Hindi Machine Translation and Language Divergence. Journal of Machine Translation, 16(4), 251-304.

Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S, 2008. "Tamil Part-of-Speech tagger based on SVMTool", In Proceedings of the COLIPS International Conference on natural language processing (IALP), Chiang Mai, Thailand. 2008. P.No:

Dhanalakshmi V, Anand kumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S. 2008. "Tamil Part-of-Speech tagger based on SVMTool", Proceedings of International Conference on Asian Language Processing 2008 (IALP 2008), November 2008, Chiang Mai, Thailand.

Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P., Rajendran S., 2009 "Chunker For Tamil Using Machine Learning" 7th International Conference on Natural Language Processing 2009 (ICON2009), IIIT Hyderabad, India.December 2009. P.No:

Dhanalakshmi.V, Anand kumar M, Rekha R U, Arun kumar C, Soman K P, Rajendran S. "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, Padmavathy P, Soman K P, Rajendran S, 2009 "Chunker For Tamil" Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom 2009, October 2009, Kottayam, India.

Dhanalakshmi. V , Anand kumar M, Soman K P, Rajendran S , 2009 "Postagger And Chunker For Tamil Language" Proceedings of TAMIL INTERNET CONFERENCE 2009 , October 2009, Cologne, Germany.

Dwivedi, S. K., & Sukhadeve, P. P. (2010). Machine Translation System in Indian Perspectives. Journal of Computer Science, 6(10), 1111-1116.
Doit, B.J. 1993. Machine translation: a view from the lexicon. Cambridge, Mass.: The MIT Press.
Anand kumar M, Dhanalakshmi V, Soman K P, Rajendran S. 2009. "A Novel Approach For Tamil Morphological Analyzer", Proceedings of TAMIL INTERNET CONFERENCE 2009, October 2009, Cologne, Germany.
Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P. and Rajendran S. (2008). "Tamil Part-of-Speech tagger based on SVMTool", In: Proceedings of the COLIPS International Conference on natural language processing(IALP), Chiang Mai, Thailand. 2008:59-64.
Dhanalakshmi V, Anandkumar M, Shivapratap G, Soman, K P and Rajendran S. (2009). Tamil POS Tagging using Linear Programming, In: International Journal of Recent Trends in Engineering, 1(2):166-169.
Dhanalakshmi V, Anand kumar M, Rajendran S and Soman K P. (2009). "POS Tagger and Chunker for Tamil Language," Proceedings of the 8th Tamil Internet Conference. Cologne, Germany, 2009.
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. and Anand Kumar M. "Hierarchal POS tagging for Tamil language using Machine learning approach", Unpublished. Uploaded in academia. edu.
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.
Dhanalakshmi V, Anand kumar M, Rajendran S, Soman K P. 2009.. POS Tagger and Chunker for Tamil Language.
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," artcom, pp.433-435, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009,IEEE Press,doi: 10.1109/ARTCom.2009.184.
Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P. and Rajendran S. .2009.. "Chunker for Tamil," artcom, pp.436-438, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009, IEEE Press,doi: 10.1109/ARTCom.2009.191.
Dhanalakshmi, V., Anand Kumar, M., Rekha, R.U., Soman, K.P., Rajendran, S. 2010. "Grammar Teaching Tools for Tamil Language", In: Technology for Education Conference (T4E 2010), pp. 85–88, India, 2010.
Dhanalakshmi V. 2010. Shallow Parser for Tamil. Thesis Submitted to Tamil University for the Award of the Degree of Doctor of Philosophy in Linguistics in Department of Linguistics, Tamil University, Thanjavur.

Dhanalakshmi V and Anand Kumar M, 2011. "Tamil shallow parser using machine leaning Approach", Tamil Internet.
Dhivya, R, Dhanalakshmi V, Anand Kumar M and Soman K.P. 2011. "Clause Boundary Identification for Tamil Language Using Dependency Parsing." SPIT/IPC 2011: 195-197.
Dirix, P., Schuurman, I., & Vandeghinste V. (2005). Metis II: Example-based machine translation using monolingual corpora - system description. In Proceedings of the 2nd Workshop on Example-Based Machine Translation, 43-50.
Doddington, G. 2002. "Automatic evaluation of machine translation quality using n-gram co-occurrence statiscs." In: HLT 2002: Human Languge Technology Conference: proceedings of the second intenational conference on human languge technology research, March 24-27, 2002. San Dieog, California: 138-145.
Dolon, B., Pinkham, J and Richardson, S.D. 2002. "MSR-MT: the Microsoft research machine translation system." In: AMTA (2002), 237-239.
Dorr, B.J. 1993. Machine translation: a view from the lexicon. Cambridge, Mass.: The MIT Press.
Eagles. Expert Advisory Group on Language Engineering Standards.
Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. In Proceedings of the Second Workshop on SMT, 220-223.
Dwivedi, S. K., & Sukhadeve, P. P. (2010). Machine Translation System in Indian Perspectives. Journal of Computer Science, 6(10), 1111-1116.
Eagles –Recommendations for the Morphosyntatic Annotaton of Corpora. Eagles Document EAG – TCWG-MACIR. Version of Mar, 1996.
EAMT. 1995. MT Summit V. Proceedings, Luxembourg, July 10-13, 1995.
EAMT. 2001. MT Summit VIII: machine translation in the information age, ed. B. Maegaard, 18-22 September 2001, Santiago de Compostela, Galicia, Spain.
EAMT. 2003. Controlled language translation: Joint conference combining the 7th international Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, May 15-17, 2003, Dublin City University, Ireland.
EAMT. 2004. Broadening horizons of machine translation and its applications. Proceedings of the Ninth EAMT workshop, 26-27 April 2004, University of Malta.
EAMT. 2005. 10th EAMT Conference, 30-31 May 2005: Practical applications of machine translation. Conference proceedings, Budapest, Hungary.
EAMT. 2006. 11th Annual Conference of the European Association for Machine Translation. Proceedings, June 19& 20, 2006, Oslo University, Norway.
Edmundson, H.P. (ed.) 1961. Proceedings of the National Symposium on Machine Translation held at the University of California, Los Angeles, February 2-5, 1960. London, etc.: Prentice-Hall.
Falkedal, K. (ed.) 1994. Proceedings of the evaluators' forum, April 21-24, 1991, Les Rasses, Vaud, Switzerland. Carouge/Genève: ISSCO.
Feigenbaum, E.A. and McCorduck, P. 1984. The fifth generation: artificial intelligence and Japan's computer challenge to the world. London: Joseph.
Francis, W.N. and Kucera, H. 1964. Manual of Information to accompany A Standard Corpous of present day edited American English. Department of Linguistics, Brown Uniersity, USA.

Garje,G.V. and Karate, G.K. 2013. Survey of Machine Translation Systems in India. International Journal on Natural Language Computing (IJNLC) vol. 2 No. 4, October, 2013 pp 47-67.
Garside, R.Leech, G. and Mcenery, A/ (Eds.) 1997. Corpus Annotation Linguistics Information from computer text Corpora.London: Longman.
Gerbig, A.1997. Lexical and Grammatical Variation in a Corpus: A computer Assisted study of Discourse on the Environment. London: Peter Long Publishing.
Gaspari, F. 2004. "Online MT services and real users' needs: an empirical usability evaluation." In: AMTA (2004),74-85.
Ghadassy, M., Henry.A., and Roseberry, R.L. (Eds) 2001. Small Corpus studies and ELT : Theory and practice Amsterdam / Philadelphia. John Benjamins.
Goodman, K. and Nirenburg, S. (eds.) 1991. The KBMT project: a case study in knowledge-based machine translation, San Mateo,CA: Morgan Kaufmann.
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. Amserdam: Rodopi.
Granger,S. Hung, J.And Tyson, S.P.Eds. 2002. Computer Learner Corpora, Second Language Acquistion and Foreign Language Teaching. Amsterdam: John Benjamins.
Greenbaun, S. (Ed.) 1996. Camparing English Worldwide: the International Corpus of English. Oxford: Clarendon.
Groves, D. & Way, A. 2005. Hybrid example-based SMT: the best of both worlds. In Proceedings of the ACL Workshop on Building and Using Parallel Texts, 183-190.
Haan, Peter de, 1992. The optimum Corpus sample size? in New directions in English Language Corpora (ed.,) Mouton de Gruyter, New York.
Hutchins, W.J. 1986. Machine Translation: past, present and future. Chichester (UK): Ellis Horwood, 1986. (ISBN: 0-85312-788-3); New York: Halsted Press, 1986. (ISBN: 0-470-20313-7).
Hutchins, W.J. 1988. "Recent developments in machine translation: a review of the last five years." In: Maxwell, D. et al (eds.) New directions in machine translation, 9-63, Dordrech: Foris.
Hutchins, W.J. and Somers, H. L. 1992. An Introduction to Machine Translation. Academic Press Limitted, London.
Hutchins, W.J. 1994. "Research methods and system designs in machine translation: a ten-year review, 1984-1994." In: Machine Translation, Ten Years on, 12-14 November 1994, Cranfield University, 16pp.
Hutchins, W.J. 1997. "From first conception to first demonstration: the nascent years of machine translation, 1947-1954. A chronology." Machine Translation 12(3), 195-252.
Hutchins, W.J. 1999. "The development and use of machine translation systems and computer-based translation tools." International Conference on Machine Translation & Computer Language Information Processing. Proceedings of conference, 26-28 June 1999. Beijing, China ed. Chen Zhaoxiong, 1-16 [Beifing: Research Centre of Computer and Languge Engineering, Chinese Acaemy of Sciences.].
Hutchins, W.J. (ed.) 2000. Early years in machine translation: memoirs and biographies of pioneers. Amsterdam/Philadelphia: Jonn Benjamins. (Studies in the Hisory of the Language Sciences, 97).

Hutchins, W.J. and Lovtsky, E. 2000. "Pert Petrovisch Troyanskii (1894-1950): A forgotten pioneer of machine translation." Machine Translation 15 (3), 187-221.
Hutchins, W.J. 2003. "ALPAC: the (in)famous report." In: Nirenburg et al (2003), 131-135.
Hutchins, W.J. 2004. "The Georgetown-IBM experiment demonstrated in January 1954." In: AMTA (2004). 102-114.
Hutchins, W.J. 2005. "Towards a definition of example-based machine translation." In: AMTA (2005), Proceedings of Second Workshop on Example-Based Machine Translation; 63-70.
Hutchins, W.J. Machine translation: a concise history. Journal of Translation Studies, vol.13, nos.1-2 (2010). Special issue: The teaching of computer-aided translation, ed. Chan Sin Wai. (Chinese University of Hong Kong, 2010); pp.29-70. [PDF, 216KB]
[Website: http://ourworld.compuserve.com/homepages/WJHutchins]
Hunston, S. 2002. Corpora in Applied Linguistics, Cambridge University Press.
Indian Language to Indian Language Machine Translation System (ILMT) System: Software Requirement Specifications version 1.03 October 2008 IIIT.
Jayaram, B.D. 1996. Development of corpora in Indian languages: Problems and suggested solutions. Paper presented at workshop on Indian Language Corpus and its applications at CIIL, Mysore.
Jayaram, B.D and Rajyashree, K.S. 1996. "Corpora in Indian Languages". http://www.emille.lancs.ac.uk/lesal/ciil.pdf
Jayaram, B.D., Umarani P. 1998. "Grammatical Category Disambiguation: A Probabilistic Model." In: South Asian Language Review, Vol. vII. No.1.
Johansson, S. and Hofland, K. (Eds) 1982. Computer corpora in English Language Research. Bergen: Nowegian computing Center for the Humanities.
JEIDA, 1992. JEIDA methodology and criteria on machine translation evaluation. Tokyo: Japan Electronic Industry Development Association.
Johansson, S. and Stenstrom, A.B. (Eds.) 1991. English computer Corpora: Selected Paper and Research Guide. Berlin: Mouton de Gruyter.
Josan G. S. and Lehal, G. S. (2008) "A Punjabi to Hindi Machine Translation System", in proceedings of COLING-2008: Companion volume: Posters and Demonstrations, Manchester, UK, pp. 157-160.
Kamakshi and Rajendran. 2008. Preliminaries to the Preparation of a Machine Aid to Translate Lingustic Texts in English into Tamil. DLA Publications, Thiruvananthapuaram.
Kay, M. .1980. "The proper place of men and machine in language translation." Research report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA, USA. Reprinted in Nirenburg et al (2003), 221-232: and (with peer group commentary)in: Machine Translation 12 (1-2), 1997:3-23.
Kennedy, G. 1998. An Introduction to Corpus Linguistics. New York: Addison-Wesley Longman Inc.
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 G.A. Rodopi.
King, M (ed.) 1987. Machine translation today: the state of the art. Edinburgh: Edinburgh University Press.
King, M., Popescu-Belis, A. and Hovy, E. 2003. "FEMTI: creating and using a framework for MT evaluation." In: AMTA (2003), 224-231.

Kirk, J.M. (Ed.) 2000. Corpora Galore: Analysis and Techniques in Describing English, Amserdam, Atlanda, G.A. Rodopi.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Koehn, P. & Hoang, H. (2007). Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods. In NLP and Computational Natural Language Learning, 868-876.

Kumar, R. and S. Viswanathan. 2001 "Tamil Morphological Analyser" at http:/www.av. KBC.org research-areas/projects/documentation.html (17.9.2001)

Kurematsu, A. and Morimoto, T. 1996. Automatic speech translation: fundamental technology for future cross-language communication. Amsterdam: Gordon and Breach.

Kyto, M. Thalainen, O.and Rissanen, M. (Eds.) 1998. Corpus Linguitics, hard and Soft : Proceedings of the 8 th International Conference on English Language Research on Computerised Corpora. Amsterdam: Rodopi.

Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. A. (2018). Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Latha Gore and Nishigandha Pail. 2002. "Enlish to Hindi –Traslation System". In Proceedings of Symposium on Translation Systems.IIT Kanpur. Pp 178-184.

Lancushine, I.E Carol and C.F. Meyer (Eds.) 1997. Synchronic Corpus Linguistics. Bergen, Norway: ICAME.

Lehmann. 1989. the Grammar of Modern Tamil. Pondicherry Institute of Linguistics and Culture. Pondicherry.

Levin, L. et al 2000. "The JUNUS-III translation system:speech-to-speech translation in multiple domains." Machine Translation 15 (1-2) 3-25.

Ljung, M. (Ed.) 1997. Corpus-Based Studies in English. Papers from the 17 th International Conference on English Language Research Based on Computerized Corpora. Amsterdam:Rodopi.

Locke,W.N. & Booth,A.D. (eds.) 1955. Machine translation of languages: fourteen essays.Cambridge,Mass.: Technology Press of the Massachusetts Institute of Technology.

Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206.

Mathhews, P.H. 1972. Inflectional Morphology. Cambridge University Press, London.

Mair, C. and hundt, M. (Eds.) 2000. Corpus Linguistics and Linguistics Theory. Amsterdam-Atlanda, GA: Rodopi.

Menon, A. G., S. Saravanan, R. Loganathan, and K. P. Soman, (2009). "Amrita Morph Analyzer and Generator for Tamil:A Rule-Based Approach." Proceedings of Tamil Internet Conference 2009, Cologne, Germany, October 2009.

Menon, V.K., Rajendran S Anand Kumar M and Soman K P. A new TAG Formalism for Tamil and Parser Analytics. Uploaded in academia. edu and Research Gate.

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.

Mecnery, T. and Wilson, A. 1996. Corpus Linguistics. Edinburg: Edinburgh University Press.
Mcenery, T. Rayson, P.and Wilson, A. (Eds.) 2002. A Rainbow of Corpora: Corpus Linguistics and the Language of the World. Munchen: Licolm Europe.
Meyer, C.F. 2002 English Corpus Linguistics. Cambridge University Press.
Muraki, K. and Ichiyama, S. 1982 An Overview of Machine Translation Project at NEC Corporation. Technical note. C & C Systems Research Laboratories, NEC Corporation
Nargo,M. 1984. "A framework of a machine translation between Japanese and English by analogy principle." In: Artificial and human intelligence: ed. A. Elithron & R. Banerji (Amsterdam:North Holland), 173-180. Reprinted in: Nirenburg et al (2003), 351-353.
Naskar, S., & Bandyopadhyay, S. (2005). Use of Machine Translation in India: Current Status. In Proceedings of MT SUMMIT X; September 13-15, 2005, Phuket, Thailand.
Nelson, G., Wallis, S. and Aarts, B. 2002 Exploring Natural Language: Working With the British component of the International Corpus of English. Amsterdam/Philadelphic John Benjamins.
Ney, H. 2005. "One decade of statistical machine translation." In: AMTA (2005), i-12-17.
On Knowledge-Based Machine Translation
Nirenburg, Sergei Victor Raskin, Allen Tucker. "On Knowledge-Based Machine Translation." Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics. https://www.aclweb.org/anthology/C86-1148
Nirenburg, S. (ed.). 1987. Machine translation: theoretical and methodological issues. Cambridge: Cambridge University Press.
Nirenburd, S., Somers, H. and Wilks, Y. (eds.) Readings in machine translation. Cambridge, Mass.: MIT Press.
Nomura, H. and Shimazu, A. 1982 Machine Translation in Japan. Technical note. Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation, Tokyo.
Nomura, H.; Shimazu, A.; lida, H.; Katagiri, Y.; Saito, Y.; Naito, S.; Ogura, K.; Yokoo, A.; and Mikami, M. 1982 Introduction to LUTE (Language Understander, Translator & Editor). Technical note.
Noone, G. 2003. Machine Translation - A Transfer Approach, A project report, www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/nooneg0203.pdf.Musashino Electrical Communication Laboratory, Research Division, Nippon Telegraph and Telephone Public Corporation, Tokyo.
Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosen, V., & Flickinger, D. 2007. Towards hybrid quality-oriented machine translation on linguistics and probabilities in MT. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in
Machine Translation, 144-153.
Ookes, M.p. 1998. Statistics for Corpus linguistics. Edinburgh: Edinburg University Press.
Ooi, V.B.V.1997. Computer Corpus Lexicogrphy. Edinburgh: Edinburg University Press.
Oostdijk, N.and Dehann, P. (Eds.) 1994. Corpus Based Research into Language Amsterdam-Atlanda, GA: Rodopi.
Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2002. "BLEU: a method for automatic evaluation of machine translation." In: ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, July 2002; 311-318.

Parameswari K, Sreenivasulu N.V., Uma Maheshwar Rao G & Christopher M, 2012. "Development of Telugu-Tamil Bidirectional Machine Translation System: A special focus on case divergence", in proceedings of 11th International Tamil Internet conference, pp 180-191

Partington, A. 1998. Patterns and meanings-Using Corpora for English LanguageReseach and Teaching. Amsterdam, Philadelphia: John BEnjamins.

Percy, C., Meyer, C.F., and Lancashire, I. (Eds.) 1996. Synchronic Corpus Linguistics, Amsterdam – Atlanta, G.A: Rodopi.

Peters, B.P., Collins, P. and Smith, A. (Eds.) 2002. New Frontiers of Corpus Research Language and Computers. Amsterdam – Atlanta, G.A. Rodopi.

Poornima, C., Dhanalakshmi, V., Kumar M. A., & Soman, K. P. 2011. Rule-based Sentence Simplification for English to Tamil Machine Translation System. International Journal of Computer Applications (0975 - 8887), 25(8), 38-42.

Rajendran, S. 1995. "A Machine Bridge for Tamil." Proceedings of Eighth World Tamil Conference. Thanjavur.

Rajendran, S. 1997. "Grammatical Formalism and Computational Analysis of Nominal Compounds in Tamil." South Asian Language Review, vol 7, no.1., 27-46.

Rajendran.S. 1999. "Spell and Grammer Checker for Tamil." Paper read in 27th All Indian Conference of Dravidian Lingustics.

Rajendran.S. 2000. "Preliminaries to the Preparation of spell and Grammer checker for Tamil." Language in Indian WWW.languageindia.com.

Rajendran.S. 2006. "Parsing in Tamil: Present state of Art." Indian Linguistics: 67:159 – 67.

Rajendran, S., Arulmozi, S.Ramesh Kumar and S.Viswanathan. 2003. "Computational Morphology of Verbal Complex". In B.Ramakrishna Reddy (ed.) Word Structure in Dravidian. Kuppam: Dravidian University 376 – 398.

Rajendran, S. et al. 2003. "Computational Morphology of Verbal Complex." In: B. Ramakrishna Reddy (edited) Word Structure in Dravidian, Kuppam: Dravidian University, & Language in India 3:4, www.languageinindia.com, April 2003.

Rajendran, S. (2006). Parsing in Tamil: Present state of art. Language in India 6:8, August 2006, www.languageinindia.com.

Rajendran, S. 2006. Parsing in Tamil: Present state of Art. Indian Linguistics vol. 67, 2006, 159–67.

Rajendran, s. 2006. A Survey Of The State Of The Art In Tamil Language Technology, Language in India 6:10, October 2006, www.languageinindia.com.

Rajendran, S. 2006. A Survey Of The State Of The Art In Tamil Language Technology, Language in India 6:10, October 2006, www.languageinindia.com.

Rajendran, S. (2007). Complexity of Tamil in POS Tagging. Language in India 7:1, January 2007, www.langugeinindia.com.

Rajendran, S. 2009. Dravidian WordNet. In: Proceedings of Tamil Internet Conference 2009. Cologne, Germany, October, 2009

Rajendran, S. 2010. Tamil WordNet. In: Proceedings of the Global WordNet Conference (GWC 10) 2010, IIT, Bombay.

Rajendran S., Shivapratap G., Dhanalakshmi V. and Soman K.P. (2010). Building a WordNet for Dravidian Languages. In: Proceedings of the Global WordNet Conference (GWC 10), 2010, IIT Bombay.
Rajendran, S. 2012. "Preliminaries To The Preparation Of A Spell And Grammar Checker For TamiL" Upoaded in academia.edu and Reseach Gate.
Rajendran, S., M. Anandkumar and Soman K.P. 2013. Computational approach to Word Sense disambiguation in Tamil. In: 12th International Tamil Internet Conference 2013, University of Malaya, Kuala Lumpur, Mayasia.
Rajendran S and Vasuki G. (2013). English To Tamil Machine Translation System Using Parallel Corpus. Book Uploaded in academia.edu and Research Gate.
Rajendran S. (2014). Resolution of Lexical Ambiguity in Tamil. Language in India (e-journal), 14:1, January, 2014.
Rajendran S and Anandkumrar M. (2014). "Corpus based approach for resolving verbal polysemy in Tamil". In: Proceedings of 13th International Conference on Tamil Computing and Tamil Internet (Tamil Internet 2014) from 19th to 21st September 2014 at Pondicherry University.
Rajendran S and Anandkumrar M. (2014). "Corpus based approach for resolving verbal polysemy in Tamil". In: Proceedings of 13th International Conference on Tamil Computing and Tamil Internet (Tamil Internet 2014) from 19th to 21st September 2014 at Pondicherry University.
Rajendran, S. Anand Kumar M. and Soman K.P. (2015). "Building hierarchies and networks from MRDs of Tamil". In: Proceedings of 14th International Conference on Tamil Computing and Tamil Internet (Tamil Internet 2015), at Singapore (SIM University, Singapore campus) from 30th, 31st May to 1st June 2015.
Rajendran, S. (2016). "Tamil Thesaurus to Tamil wordNet." In: Proceedings of 15th World Tamil Internet conference 2016, held at Gandhi Gram Rural University, Tamil Nadu, September 8-11, 2016, pages 1-9.
Rajendran, S. and Anandkumar, M. (2017). "Visual Onto-Thesaurus for Tamil." Language in India www.languageinindia.com ISSN 1930-2940 vol. 17:5 May 2017.
Rajendran S. (2018) "Lexical Resource Tool for Tamil Computing." In: Conferene Papers - 17th Tamil Internet Conference at Tamil Agricultural University, Coimpator.
Rajendran S and Anand Kumar (2018) "Computing tools for Tamil Language Teaching and Learning." In: Conferene Papers - 17th Tamil Internet Conference at Tamil Agricultural University, Coimpator.
Rajendran S. and Vasuki G. 2019. English Tamil Machine Translation System using parallel corpus. Language in India, www.languageinindia.com, ISSN 1930-2940 vol 19:5 May 2019.
Rao, M. D. (2000). Machine Translation in India: A Brief Survey.
www.elda.org/en/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf.
Ramaswamy, V. 2002. Morphological Generator for Tamil. Unpublished M.Phil Dissertation, University of Hyderabad, Hyderabad.
Ranganathan, 1997. "A Lexical phonology Approach Tamil word by Computer". International Journal of Dravidian Linguistics, 26, 1:57 – 70.
Renganathan, V. 2002. An Interactive Approach to Development of English-Tamil Machine Translation System on the Web. Tamil Internet 2002, California, USA. 68-73.
www.infitt.org/ti2002/hubs/ conference/papers.html.

Rekha, R U Anand kumar M, Dhanalakshmi.V , Soman K P, Rajendran S, 2010 "Morphological generator for Tamil a new data driven approach", Tamil Internet Conference 2010, June 2010, Cemmozhi maanaadu, Coimbatore, India.
Rosetta, M.T. 1994. Compositional translation. Dordercht: Kulwer.
Rozencvejg, V.Ju. (ed.) 1974. Machine translation and applied linguistics. 2 vols. Frankfurt a.M.: Athenaion Vlg. [also published as: Essays on lexical semantics, 2 vols. Stockholm: Skriptor.]
Sachi Dave, Jignashu Parikh, Pushpak Bhattacharyya. 2001. "Inteligua-based English-Hindi Machine Translation and Language Divergence." Machine Translation, 2001. Volume 16, Issue 4, pp 251-304.
Sadler, V. 1989. Working with analogical semantics: disambiguation techniques in DLT. Dordrecht: Foris.
Sampark: Machine Translation System among Indian languages (2009) http://tdildc.
in/index.php?option=com_vertical&parentid=74, http://sampark.iiit.ac.in/
Sanjay Chatterji, Devshri Roy, Sudeshna Sarkar & Anupam Basu, (2009) "A Hybrid Approach for Bengali to Hindi Machine Translation", In proceedings of ICON-2009, 7th International Conference on Natural Language Processing, pp. 83-91.
Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve. 2010. "Machine Translation System in Indian Perspective." Journal of Computer Science vol. 6(10), pp 1082, 2010, Science Publications.
Sawai, S.; Fukushima, H.; Sugimoto, M.; and Ukai, N. 1982 Knowledge Representation and Machine Translation. Proceedings of the Ninth ICCL (COLING 82), Prague, Czechoslovakia: 351-356.
Selkairk, E.O. 1983. The Syntax of words. Cambridge, Mass: MIT Press.
Selting, M.and Couper. Kuhlen E. (Eds.) 2001. Studies in International Linguistics. Amsterdam / Philadelphia: John Benjamins.
Sengupta, G. 1997. "Three models of Morphological processing" South Asian Language Review, 8:1, 1-26.
Shanmugam, C. 2001. "Computer Analysis of Simple sentence in Tamil". Paper read in UGC SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, AnnamalaiNagar.
Shannon, C.E. and Weaver, W. 1949. The mathematical theory of communication. Urbana: University of Illinois Press.
Shanmugam, C. 2002. "Grammer and Parser: A program for Syntactic Parsing in Tamil" International Seminar on Tamil Computing 27 -28 February and March 1, 2002. University of Madras: Chennai.
Sharman, R. 1990. Hidden Markov Model Methods for word Tagging. Report 214. Winchester : IBM UK Scientific centre.
Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007). Rule-based translation with statistical phrase-based post-editing. In Proceedings of the Second Workshop on SMT, 203-206.
Sinha, R.M.K. 2004. An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS-2004), Novermber 17-19, 2004.
Sinha, R.M.K. et al. 1995. ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi. IEEE International Conference on Systems, Man and Cybernetics,Vancouver, Canadam 1995, pp 1609-1614.

Sinha, R.M.K. and Jain, A. 2003. AnglaHindi: an English to Hindi machine-aided translation system, MT Summit IX, New Orleans,USA, 23-27 September 2003, pp 494-497.
Sivashanmugam, C. 1998. "Language Technology and Morphological Processing". The Administrator, 45:3.
Sitender & Seema Bawa, (2012) "Survey of Indian Machine Translation Systems", International Journal Computer Science and Technolgy, Vol. 3, Issue 1, pp. 286-290, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)
Sivashanmugam, C. 2000. "A Model for computer Analysis of Verb in Tamil" In working papers in Linguistics, Department of Linguistics, Bharathiyar University, Coimbatore.
Sivashanmugam, C. 2002. "Morphological: processor for Negative Constructions in Tamil" Indian Conference on Natural Language processing, Anna University, Chennai.
Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford University Press.
Slocum, J. 1985. A survey of machine translaton: its history, current status, and future prospects. Computational Linguistics, volume 11, number 1.
Solcum, J. (ed.)1988. Machine translation system. Cambridge: Cambridge University Press.
Somers, H.L1999 "Exampl-based machine translation." Machine Translation 14(2):113-158.
Souter, C. and Atwell, E. (Eds.) 1993, Corpus based computational Linguistics. Amsterdam: Rodopi.
Stenstorm, A.B. Anderson, G.and Hasund, I.K.2002. Trends in Teenage Talk: Corpus Compilation, Analysis and Findings, Amsterdam: John Benjamins.
Stubbs, M. 1996. Text and corpus Analysis : Computer Assisted Studies of Lanaguage and culture. Oxford : Balckwell Publishers.
Sudip Nakar and Sivaji Bandopadhyaya. 2005. "Use of Machine Translation in India: Current Status," AAMT Journal, pp 25-31.
Sugata Sanyal & Rajdeep Borgohain. 2013. "Machine Translation Systems in India", Cornel University Library, arxiv.org/ftp/arxiv/papers/1304/1304.7728.pdf
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).
Summers, D. 1991. Longman/Lancaster English Language Corpus: criteria and Design Harlow: Longman.
Svartvik, J, (Ed.) 1990. The London Corpus of spoken English: Description and Research. Lund Studies in English 82, Lund: Lund University Press.
Starvik, J. Ed. 1992. Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 – Stockholm, 4-8 August 1991. Berling, New York, Mouton De Gruyer.
Sushil Kumar and Parvin Akhter. 2015. Indian Machine Translation Systems. ISSN (print): 2393-8374, (online): 2394-0697, Volume-2, Issue-9, 2015
Thomas, J. and Short, M. (Eds.) 1996. Using Corpora for Language Research: Studies in the Honour of Geoffery Leech. London and New York: Addison Welsely Longman.
Tognini–Boneli. E.2001. Corpus Linguistics at work. Amsterdam: John Benjammins.
TMI. 1988. Second International Conference on Theoretical and Methodological Issues in Machine Translation of
Natural Languages, June 12-14, 1988. Pittsburgh, PA, USA: Carnegie Mellon University, Center for

Machine Translation.
TMI. 1990. The Third International Conference on Theoretical and Methodological Issues in Machine Translation
of Natural Languages, 11-13 June 1990. Austin, TX: Linguistics Research Center, University of Texas.
TMI. 1992. Quatrième Colloque international sur les aspects théoriques et méthodologiques de la traduction
automatique. Fourth International Conference on Theoretical and Methodological Issues in Machine
Translation: Empiricist vs. rationalist methods in MT. Actes du colloque. Proceedings of the Conference, June 25-27, 1992, Montréal, Canada.
TMI. 1993. TMI-93: The Fifth International Conference on Theoretical and Methodological Issues in Machine
Translation with special emphasis on: MT in the Next Generation, Kyoto International Community
House, Kyoto, Japan, July 14-16, 1993.1995. The Sixth International Conference on Theoretical and Methodological Issues in Machine
Translation, TMI-95, July 5-7 1995, Katholieke Universiteit Leuven, Centre for Computational
Linguistics, Leuven, Belgium
TMI. 1997. TMI-97:Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, July 23-25, 1997, St. John's College, Santa Fe, New Mexico USA
TMI. 1999. Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99), August 23-25, 1999, University College, Chester, England.
TMI. 2002. Proceedings of the 9th International Conference on Theoretical and Methodological issues in Machine Translation, March 13-17, 2002, Keihanna, Japan
TMI. 2004. TMI-2004: Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, October 4-6, 2004, Baltimore, Maryland, USA
Umarani, P. 2001. Morphological Analyser for Tamil. Unpublished Doctoral Dissertation.
Unnikrishnan, P., Antony, P. J., & Soman, K. P. (2010). A Novel Approach for English to South Dravidian Language SMT System. International Journal on Computer Science and Engineering (IJCSE), 02(08), 2749-2759.
Vishal Goyal, Gurpreet Singh Lehal, 2011. "Hindi to Punjabi Machine Translation System." In the Proceedings International Conference for Information Systems for Indian Langauges, Patiala, Department of Computer Science, Punjabi University, Patiala, March 9-11, 2011, pp 239-241, Springer CCIS 139. Germany.
Vishal Goyal & Gurpreet Singh Lehal, (2011) "Hindi to Punjabi Machine Translation System", in proceedings of the ACL-HLT 2011 System Demonstrations, pages 1–6, Portland, Oregon, USA, 21 June 2011.
Vaishnavi Ramaswamy. 2000. A Morphological Generator for Tamil. M.Phil. Thesis, CALTS. University of Hyderabad, Hyderabad.
Vasconcellos. M. (ed.) 1994. MT evaluation: basis for future directons. Proceedings of workshop…2-3 November 1992. Washington, DC: Association for Machine Translation in the Americas.
Veronis, T. (Ed.) 2000. Parallel Text Processing. Alignment and use of Translation Corpors. Dordreeht: Kluwer Academic Publishers.

Vijayanand K., Choudhury S.I., Ratna P.2002. VAASAANUBAADA- Automatic Machine Translation of Bilingual Bengali-Assamese New Texts. Language Engineering Conference, 2002, Hyderabad, India, pp 183-188.

Wahksterm W. (ed.). Verbmobil: foundation of speech-to-speech translation. Berlin: Springer.

Whitelock, P. and Kilby, K. 1995. Linguisic and Computational techniques in machine translation system design.2nd ed.London: UCL Press.

Wahlster, W. (ed.) 2000. Verbmobil: foundations of speech-to-speech translation. Berlin: Springer.

Whitelock, P. and Kilby, K. 1995. Linguistic and computational techniques in machine translation system design. 2nd ed. London: UCL Press. [First published in 1982 as an internal report of the Centre for Computational Linguistics, University of Manchester Institute of Science and Technology, Manchester, UK.]

Wichmann, A. Fligelstone, S. Mcenery, T. and Knowles, G (Eds.) 1997. Teaching and Language Corpora, London: Longman.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Young, S. and G.Bloothooft (Eds.) 1997. Corpus based methods in Language and Speech Processing, Vol. II, Dordrecht: Kluwer Academic Press.

Zhang, Y. (2006). Chinese-English SMT by Parsing. www.cl.cam.ac.uk/~yz360/mscthesis.pdf.