==========
Language in India www.languageinindia.com **ISSN 1930-2940 Vol. 19:9 September 2019**
==========

Word Sense Disambiguation in Tamil

Prof. Rajendran Sankaravelayuthan
Amrita Vishwa Vidyapeetham, Coimbatore
rajushush@gmail.com
&
Dr. A. Dhanavalli
Research Assistant
Department of Linguistics, Tamil University, Thanjavur
danavalliraja@gmail.com

==========

ABSTRACT

This monograph on "Word Sense Disambiguation in Tamil" is organized into six chapters: 1. Introduction, 2. Structural ambiguity and lexical ambiguity, 3. Machine translation and word sense disambiguation, 4. Word sense disambiguation processing and development, 5. Approaches to word sense disambiguation in Tamil, and 6. Conclusion.

Chapter 1: Introduction

The introduction deals with aims and objectives and the importance of word sense disambiguation in machine translation.

Chapter 2: Structural ambiguity and lexical ambiguity

There are two kinds of ambiguity: structural ambiguity and lexical ambiguity. If a sentence or phrase has more than one interpretation of meaning, it is called structural ambiguity. For example, the following two phrases can be interpreted to have at least two meanings.

veLLai marundtu kuppi

'the white bottle having medicine'

'the bottle having white medicine

pazaiya maaNavar viTuti

'the hostel which belongs to old students'

'the old hostel which belongs to students'

In the following examples, two interpretation of meaning is possible as the word *kaal* has two meanings: 'one fourth' and 'leg'.

avan kooziyin kaal pakutiyaic caappiTTaan

'he ate quarter part of a chicken'

'he ate the leg part of the chicken'

There are a few types of lexical ambiguity: categorical ambiguity, ambiguity due to homonymy and ambiguity due to polysemy. The word *meelee*, for example, belongs to three grammatical categories as shown below.

avan meelee irukkiRaan 'He is at the top'

avan meelee cenRaan 'He went up'

avan meejai meelee eeRinaan 'He climbed on the table'

The word *meelee* in the first example belongs to noun, *meelee* in the second example belongs to adverb and *meelee* in the third example belongs to postposition. So *meelee* is potential enough to show categorical ambiguity in three ways. Similarly the word form *ndey* belongs to two categories as it can denote the noun *ndey* 'ghee' and the verb *ndey* 'wave'.

avan tuNi ndeytaan 'He weaved a cloth'

avaL ndeytaan virumpukiRaaL 'She likes only ghee'

In the absence of a context, *meelee* and *ndey* are cateogically ambiguous. For a machine *meelee* and *ndey* are always categorically ambiguous as it is not given the knowledge a human being posseses.

In the following example, ambiguity arises due to homonymy.

kuzantdai aTTaiyaip paarttu cirittatu 'The child laughed at the leech/wrapper'

avan puttakattin aTTaiyaik kizittaan 'He tore the wrapper of the book'

avan aTTaiyaik kampaal aTittuk konRaan 'He killed the leech with a stick'

The first example can be interpreted in two ways as *aTTai* can mean 'leech' or 'wrapper'. In the second and third example, the ambiguity is resolved by context; in the second example, *aTTai* means 'wrapper' and in the third example it means 'leech'.

Chapter 3: Machine Translation and word sense disambiguation

This chapter explains word sense disambiguation in the context of machine translation. This chapter is based on "Words Sense Disambiguation: The state of Art" written by Nancy Ide and Jean Veronis (1998).

Word sense disambiguation is useful for the following applications:

- Machine translation
- information retrieval and hyper text navigation
- content and thematic analysis
- grammatical analysis
- speech processing and
- text processing

Automated sense disambiguation was attempted in the context of machine translation. Weaver talks about the importance of sense disambiguation in his 'memorandum'. Kaplan (1950) tried to explain about sense disambiguation by an experiment. Reifler (1955) pointed out the importance of context in terms of "semantic coincidence". Consecutively semantic representation, interlingual device and semantic network came into vogue.

1960s relied upon AI based methods. Symbolic methods, connectionist methods and knowledge based methods were attempted. In eighties, machine readable dictionaries came into vogue. Lesk (1989) developed knowledge base for each dictionary sense. Thesauri have also been used for resolving polysemy. Yerowsky (1992) developed some methodology in this line of thinking. In the midst of 1980s, computational lexicons came into vogue. The multilingual on line lexical data bases like wordNets have been built in 1990s. But wordNet are found to be not very useful for resolving word sense ambiguity. Pustejovsky elaborated on

generative lexicon (Pistejovsky 1995 and Pustejovksy et al 1905) and pointed out the need to generate meanings form event structure, qualia structure and lexical inheritance structure.

The last phase of 19th century and the beginning of 20th century saw the development of corpus based approaches. Weiss (1973) and Kelley and Stone (1975) proved that WSD can be performed by making use of corpus. Attempts were made for automatic sense-tagging. Brown et al (Brown et al 1991) and Gale et al proposed the use of parallel corpus for selecting words across languages. Methods were proposed for overcoming scarcity in the parallel corpus. Smoothing, class-based models and similarity based models were proposed.

The role of context in sense selection is widely felt. The context for WSD has been manipulated specifically in two approaches: bag-of-words approach and syntactic relation approach. The contexts such as microcontext, topical context and domain specific contexts were recognized. Distance, collocation, and syntactic relations have been taken into account. The issues such as 'senses or usages' and 'enumeration or generation' have been raised.

Chapter 4: Process of Word Sense Disambiguation and its development

Word sense disambiguation has been described as AI-complete problem. There are approaches such as lexical sample or targeted WSD and All-words WSD. There are two types of resources, internal resources, unstructured resources, raw sources, sense-annotated corpus and external resources. The external sauces consists of thesaurus, machine readable dictionaries, ontologies, collocational resources and the other sources such as wordNet,

There are at least three types of approaches: supervised approaches unsupervised approaches and knowledge-based approaches. Supervised includes the following: decision lists, decision trees, Naïve Bayes, neural network, example-based or instance based learning, Support Vector Machines and semi supervised WSD. The supervised methods include context clustering, word clustering and co-occurring graphs. Knowledge-based approach includes over lapping of sense definition and graph based approaches. The other approaches include determining word sense dominance, domain-driven disambiguation, and WSD from Cross-Lingual Evidence, The uses of WSD include information retrieval (IR), information extraction,

machine translation, content analysis, word processing, lexicography and semantic web. The problem includes semantic representation, knowledge acquisition and domain-driven disambiguation.

Chapter 5: Word sense disambiguation processing in Tamil

To start with Tamil lexical semantic is discussed in details. Different types of lexical relations such as synonymy, hyponymy, compatibility, incompatibility, meronymy and antonymy are discussed with suitable examples. The vocabulary of Tamil is structured hierarchically by taxonomy and meronymy relations. The structures could be branching or non-branching.

There are few attempts for word sense disambiguation in Tamil. The first attempt of WSD in Tamil has been made by Arun. This is followed by Baskaran's research on word sense disambiguation (Baskaran 2002). He has made use of cluster approach complemented by case information. In this work, disambiguation is being performed based on the clustering technique and by using the collocations and case-markers as knowledge sources. The approach works in two phases: trainging phase and testing phase. At the beginning of the training phase, the system identifies the ambiguous words in the text using the list of ambiguous words. For each occurrence of an ambiguous word a context is identified from the text. Different researchers have used different sizes of contexts, i.e. different window sizes. The window size influences the performance of the system considerably. A larger window size brings in more unrelated words to the context, while a smaller window size misses some important collocations. Following the general practice as found in many WSD works including Yarowsky (1995), a twenty-word window is used in this work. The frequent words also called stop-words are removed from the window. The context is then morphologically analysed to get the root form of the contextual words, which are used, in clustering. Each context is represented as a context or co-occurrence vector. Using the K-means approach these vectors are collected into different clusters. Collocations are then collected from these clusters automatically and senses are assigned to these collocations by human-annotators thereby developing a sense collocation dictionary.

Case-markers are also identified for each cluster and added to the sense-collocation dictionary to build a sense-collocation case-marker dictionary.

In the second phase, the system is ready for disambiguation. Any text containing an ambiguous word can be given to the system and the system makes use of the sense-collocation case-marker dictionary to disambiguate the unseen occurrences of the ambiguous word. It should be noted that the training is a one-time task for each ambiguous word and once the system is trained for a particular word it can be used to resolve ambiguity of that word in a given text. It can be observed that collocations and case-markers play an important role in this method.

A methodology of making use of conceptual graph for WSD has been explained next to it. This is followed by making use of Rajendran's computational Thesaurus for Tamil (2006) for WSD.

Chapter 6: Conclusion

The last chapter is “Conclusion”. It discusses briefly about the information discussed in the last five chapters. It also discusses about the contribution made by the research work.

Keywords: word sense disambiguation, lexical ambiguity, structural ambiguity, ambiguity due to homonymy, ambiguity due polysemy, categoriacal ambiguity, transfer ambiguity, contextual knowledge, referential ambiguity, scope ambiguity, automated sense disambiguation, sense disambiguation component, connectionist method, knowledge based methods, machine readable dictionaries, computational lexicon, enumerative lexicon, generative lexicon, automatic sense tagging, enumeration, generation, natural language processing, knowledge extraction, all-words WSD, enumerative approach, ontology, unstructured resources, corpora, collocational resources, wordNet, synonymy, hyponymy, meronymy, hypernymy, classification method, unsupervised WSD, supervised WSD, knowledge based WSD, decision list, decision tree, Naïve Bayes, neural network, support vector machine, ensemble method, Adaboost, bootstrapping, word clustering, graph-based approach, co-occurrence graph, knowledge based disambiguation, similarity measures, domain-driven disambiguation, compatibility,

incompatibility, antonymy, vector space, clustering, conceptual graph, canonical graph, pragmatic context.

தமிழில் சொற்பொருள்மயக்கநீக்கம்

(Word sense disambiguation in Tamil)

ஆக்கியோர்

பேராசிரியர் இராசேந்திரன் சங்கரவேலாயுதன்

அமிர்தா விஷ்வ விதயபீடம், கோயம்புத்தூர்

rajushush@gmail.com

&

முனைவர் அ. தனவள்ளி

ஆய்வு உதவியாளர், மொழியியல்துறை, தஞ்சாவூர்

dhanavalliraja@gamil.com

கோயம்புத்தூர்

செப்டம்பர், 2019

பொருளடக்கம்

வரிசை எண்	தலைப்பு	பக்கம்
1	**இயல் 1:அறிமுகம்**	16
2	**இயல் 2: அமைப்புப் பொருண்மை மயக்கமும் சொற்பொருண்மை மயக்கமும்**	21
2.0	முன்னுரை	21
2.1	பொருண்மை மயக்கம்	21
2.2	பொருண்மை மயக்கத்தின் வகைகள்	22
2.2.1	சொற்பொருண்மை மயக்கம்	23
2.2.1.1	சொற்பொருண்மை மயக்கத்தின் வகைகள்	24
2.2.1.1.1	சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம்	24
2.2.1.1.2	ஒப்புருச்சொன்மைசார், பல்பொருள் ஒருமொழியம்சார் பொருண்மை மயக்கம்	24
2.2.1.1.3	மாற்றப் பொருண்மை மயக்கம்	30
2.2.2	அமைப்புப் பொருண்மை மயக்கம்	30
2.2.2.1	அமைப்புப் பொருண்மை மயக்கத்தின் வகைகள்	31
2.2.2.1.1	உண்மையான அமைப்புப் பொருண்மை மயக்கம்	32
2.2.2.1.2	யதேட்சை அமைப்புப் பொருண்மை மயக்கம்	33
2.2.2.2	அமைப்புப் பொருண்மை மயக்கத்தை நீக்குதல்	36
2.2.2.2.1	மொழியியல் அறிவைப் பயன்படுத்தல்	36
2.2.2.2.2	சூழல் அறிவு	38
2.2.2.2.3	உண்மையான உலக அறிவு	38
2.2.2.2.4	பிற உபாயங்கள்	39
2.2.3	குறிப்புப்பொருள் பொருண்மை மயக்கம்	40

2.2.4	நோக்கப் பொருண்மை மயக்கம்	41
2.2.5	பொருண்மை மயக்கம் வேறுபடுத்தல்	42
2.2.6	தத்துவ சம்பந்தம்	43
2.3	முடிவுரை	43
3	**இயல் 3: இயந்திர மொழிபெயர்ப்பும் சொற்பொருண்மை மயக்கநீக்கமும்**	45
3.0	முன்னுரை	45
3.1	சொற்பொருண்மை மயக்க நீக்கத்தின் கருத்தாய்வு	47
3.1.1	இயந்திர மொழிபெயர்ப்பில் தொடக்ககாலச் சொற்பொருண்மை மயக்கநீக்கம்	49
3.1.2	செயற்கை அறிவு அடிப்படையிலான நெறிமுறைகள்	53
3.1.2.1	குறியீடு சார் நெறிமுறைகள்	53
3.1.2.2.	தொடர்புபடுத்துபவர் நெறிமுறைகள்	57
3.1.3	அறிவு அடிப்படையிலான நெறிமுறைகள்	59
3.1.3.1	இயந்திரம் படிக்கவியலும் அகராதிகள்	60
3.1.3.2	பொருட்புல அகராதிகள்	61
3.1.3.3	கணினிசார் பேரகராதிகள்	62
3.1.4	தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள்	63
3.1.4.1	அனுபவ நெறிமுறைகளின் வளர்ச்சி, வீழ்ச்சி, மீட்சி	63
3.1.4.2	தானியக்க அர்த்தம் அடையாளப்படுத்தல்	64
3.1.4.3	தரவு அரிதாக இருப்பதை நேரிடுவது	66
3.2	திறந்த சிக்கல்கள்	69
3.2.1	சூழலின் பங்களிப்பு	69
3.2.1.1.	நுண்சூழல்	70
3.2.1.2	தலைப்பு சார் சூழல்	74
3.2.1.3	பொருட்புலம்	77
3.2.2	பொருண்மை பகுப்பு	78

3.2.2.1	வங்கி மாதிரி	78
3.2.2.2	நுணுக்கம்	78
3.2.2.3	பொருண்மைகளா? பயன்பாடுகளா?	80
3.2.2.4	பட்டியலிடுதலா அல்லது ஆக்குதலா?	81
3.3.	முடிவுரை	82
4	**இயல் 4: சொற்பொருண்மை மயக்கநீக்கத்தின் செயல்பாடுகளும் முன்னேற்றங்களும்**	84
4.0	முன்னுரை	84
4.0.1	நோக்கம்	84
4.0.2	சுருக்கமாக வரலாறு	88
4.0.3	சுருக்கம்	88
4.1	செயல்பாட்டு விளக்கம்	89
4.1.1	சொற்பொருண்மைகளின் தேர்வு	91
4.1.2	புற அறிவு மூலங்கள்	96
4.1.2.1	சொல்வலை	100
4.1.2.2	செம்கோர்/SemCor	105
4.1.3	சூழலின் உருப்படுத்தம்	106
4.1.4	வகைபடுத்தும் நெறிமுறைகளின் விருப்பத்தேர்வு	111
4.2	கண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம்	115
4.2.1	தீர்மானப் பட்டியல்கள்	116
4.2.2	தீர்மானக் கிளைகள்	118
4.2.3	நெய்வ் பெய்ஸ்	119
4.2.4	நரம்புசார் வலையமைப்பு	121
4.2.5	முன்மாதிரி அடிப்படையிலான அல்லது எடுத்துக்காட்டு அடிபடையிலான கற்றல்	123
4.2.6	சப்போர்ட் வெக்டர் மெஷியன்கள்	125

4.2.7	ஒருங்கிணைந்த நெறிமுறைகள்	127
4.2.7.1	பெரும்பான்மை வாக்களிப்பு	128
4.2.7.2	நிகழ்தகமை கலப்பு	128
4.2.7.3	தர அடிப்படையிலான ஒருங்கிணைப்பு	129
4.2.7.4	அடாபூஸ்ட்	129
4.2.8	மிகக்குறைவான மற்றும் பகுதிகண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம்	130
4.2.8.1	மேம்படுத்தல்	130
4.3	கண்காணிக்கப்படாத பொருண்மைமயக்கநீக்கம்	131
4.3.1	சூழல் கொத்தாக்கம்	133
4.3.2	சொல் கொத்தாக்கம்	134
4.3.3	சேர்ந்துவருகை வரைபடங்கள்	136
4.4.	அறிவு-அடிப்படையிலான பொருண்மை மயக்கநீக்கம்	138
4.4.1	பொருண்மை வரையறைவிளக்கங்களின் மேலுறல்	139
4.4.2	விருப்பத்தேர்வுகள்	140
4.4.3	அமைப்புமுறை அணுகுமுறைகள்	141
4.4.3.1	ஒற்றுமை அளவீடுகள்	142
4.4.3.2	வரைபட அடிப்படையிலான அணுகுமுறைகள்	143
4.5	பிற அணுகுமுறைகள்	145
4.5.1	சொற்பொருண்மை ஆதிக்கத்தை நிர்ணையித்தல்	145
4.5.2	பொருட்புல இயக்க பொருண்மை மயக்கநீக்கம்	145
4.5.3	சொற்பொருண்மை மயக்கநீக்கத்திலிருந்து மொழி கடந்த சான்று	146
4.6	மதிப்பீட்டு நெறிமுறை	149
4.6.1	மதிப்பீட்டு நடவடிக்கைகள்	149
4.6.2.	தொடக்கநிலைகள்	149
4.7.	மதிப்பீடு: சென்ஸ்வல்/செம்வல் போட்டிகள்	149
4.7.1	சென்ஸ்வல்	150

4.7.2	சென்ஸ்வல்/செம்வல் போடிகளைப்பற்றிய கூற்றுகள்	150
4.8.	பயன்பாடுகள்	151
4.8.1	தகவல் மீட்பு	151
4.8.2	தகவல் பிரித்தெடுப்பு	152
4.8.3	இயந்திர மொழிபெயர்ப்பு	153
4.8.4	பொருளடக்க ஆய்வு	154
4.8.5	சொல் பகுப்பாய்வு	154
4.8.6	அகராதியியல்	155
4.8.7	பொருண்மை வலை	155
4.9	திறந்த சிக்கல்களும் எதிர்காலத் திசைகளும்	155
4.9.1	சொற் பொருண்மையின் உருப்படுத்தம்	155
4.9.2	அறிவுப் பேறு நெருக்கடி	157
4.9.3	பொருட்புலம் சார் சொற்பொருண்மை மயக்கநீக்கம்	159
4.10	முடிவுரை	159
5	**இயல் 5: தமிழில் சொற்பொருள் மயக்கநீக்கத்திற்கான நெறிமுறைகள்**	162
5.0	முன்னுரை	162
5.1	தமிழ்ச் சொற்பொருண்மையியல் ஆய்வு	162
5.1.1	சொற்களின் பொருண்மை உறவுகள்	162
5.1.1.1	உறுப்பமைவு உறவு	163
5.1.1.2	அடுக்கு உறவு	163
5.1.2	அடிப்படை உறவுகள்	163
5.1.3.	சொல்லுறவுகள்	164
5.1.3.1	ஒருபொருள் பன்மொழியம்	164
5.1.3.2.	உள்ளடங்கு மொழியம்	168
5.1.3.3	இணக்கம்	171
5.1.3.4.	இணக்கமின்மை	172

5.1.3.5.	எதிர்மொழியம்	172
5.1.3.6	பகுதி-முழுமை உறவுகள்	176
5.1.4.	பிற பொருண்மை உறவுகள்	178
5.1.4.1.	பகுதி உறவுகள்	178
5.1.4.2.	முழுமையுறா உறவுகள்	179
5.1.4.3.	போலி உறவுகள்	179
5.1.5.	சொற்றொகுதியின் பொருண்மை அமைப்பு	180
5.1.5.1.	படிநிலை அமைப்பு	180
2.1.5.2.	விகிதத் தொடர்கள்	180
5.1.5.3.	சொற்றொகுதியின் படிநிலை அமைப்பு	182
5.2.	தமிழில் சொற்பொருண்மை மயக்கநீக்க முயற்சிகள்	191
5.2.1	அருணின் சொற்பொருண்மை மயக்கநீக்க முயற்சி	191
5.2.1.1	நெறிமுறை	191
5.2.1.2	சொல்மயக்கநீக்க நெறிமுறைகள்	192
5.2.1.2.1	சொற்பொருண்மை மயக்கநீக்கத்திற்கு பெய்சின் வகைப்படுத்தி	192
5.2.1.2.2	அகராதி அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் (கண்காணிப்பற்றது)	192
5.2.1.2.3	தகவல் கோட்பாடுசார் அணுகுமுறை	192
5.2.1.2.4	கண்காணிக்கப்படாத வழிமுறை வரைவு கற்றல்	192
5.2.1.2.5	வாஸ்ப்-பெஞ்	193
5.2.1.2.6	மதிப்பீடு	193
5.2.2	பாஸ்கரனின் சொற்பொருண்மை மயக்கநீக்க ஆய்வு	193
5.2.2.1	அணுகுமுறை	194
5.2.2.1.1	சேர்ந்துவருபவைகளின் முக்கியத்துவம்	196
5.2.2.1.2	வேற்றுமை குறியீடுகளின் முக்கியத்துவம்	197
5.2.2.2	கற்றல் கட்டம்	198
5.2.2.2.1	சூழல்களும் சூழல் இடைவெளியும்	198

5.2.2.2.2	சூழல் இடைவெளியில் சூழல்களை உருப்படுத்தம் செய்தல்	199
5.2.2.3	பரிசோதனைக் கட்டம்	203
5.3	கருத்துரு வரைபட அகராதி அடிப்படையில் சொற்பொருண்மை மயக்க நீக்கம்	206
5.3.1	கருத்துரு வரைபடம்	206
5.3.1.1	வாய்பாட்டுக் கருத்துரு வரைபடம்	208
5.3.1.2	பொதுமையாக்கமும் சிறப்பாக்கமும்	208
5.3.1.3	சுருக்கமும் வரையறை விளக்கமும்	208
5.3.1.4	ஒன்றுதிரட்டலும் தனிநிலைப்படுத்தலும்	209
5.2.1.5	திட்டவரைவும் முன்மாதிரியும்	210
5.2.2	கருத்துரு வரைபட அகராதி உருவாக்கத்திறகான மூலவளங்கள்	212
5.2.2.1.	இராசேந்திரனின் மின்சொற்களங்சியம்	212
5.2.2.2	கிரியாவின் தற்காலத்தமிழ் அகராதி	222
5.2.3	கருத்துரு வரைபட அகராதியைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்தல்	222
5.2.4	முடிவுரை	233
6	**இயல் 6: இறுதியுரை**	235
6.1	ஆய்வுச்சுருக்கம்	235
6.2	ஆய்விலிருந்து அறியப்பட்ட உண்மைகளும் கண்டுபிடிப்புகளும்	238
6.3	ஆய்வின் பயன்பாடுகள்	240
6.4	எதிர்கால நடவடிக்கைகளும்	240
	துணை நுற்பட்டியல்	242

இயல் 1

அறிமுகம்

ஒரு சொல்லுக்குப் பல அர்த்தங்கள் வருவதை நாம் அறிவோம். ஒரு குறிப்பிட்ட சூழலில் ஒரு சொல்லைப் பயன்படுத்துவதன் மூலம் அச்சொல்லின் எந்த "அர்த்தம்" (பொருள்) ஊக்குவிக்கப்படுகிறது என்பதை தீர்மானிப்பதில் சிக்கல் உள்ளது. இச்சிக்கல் பெரும்பாலும் சாதாரண மக்களால் உணரப்படுவதில்லை. இயற்கை மொழி ஆய்வுச் செயல்பாடுகளில் இது ஒரு சவாலாக உள்ளது. மனித மூளை சொற்பொருண்மை மயக்கத்தை நேரிடுவதிலும் அதை நீக்கிச் சரியான அர்த்தத்தைப் புரிந்துகொள்வதிலும் மிகவும் திறமையாகச் செயல்படுகின்றது. இயற்கை மொழி பொருண்மை மயக்கத்தை நேரிடும் வகையில் உருவாக்கப்படுள்ளது. இது மனித உடலின் நரம்பியல் அமைப்பின் பிரதிபலிப்பாகும். வேறு வார்த்தைகளில் கூறுவதானால், மூளையின் நரம்பியல் வலைப்பின்னல்கள் வழங்கிய உள்ளார்ந்த திறனை பிரதிபலிக்கும் வகையில் மனித மொழி வளர்ந்துள்ளது (வடிவமைக்கப்பட்டுள்ளது).

மனித மூளை சொல்வரும் சூழலைப் பயன்படுத்தி சரியான அர்த்தத்தைத் தேர்ந்தெடுத்துப் புரிந்து கொள்கின்றது. ஆனால் கணிப்பொறிக்கு மனிதன் மூளையின் திறன் இல்லை. மனித மூளையின் இத்திறன் கணிபொறிக்கு வழங்கப்பட்டால் இது சாத்தியமாகும். கணினி மொழியியலில், சொற்பொருண்மை மயக்கநீக்கம் (Word Sense Disambiguation (WSD)) என்பது ஒரு வாக்கியத்தில் ஒரு சொல் எந்த அர்த்ததில் பயன்படுத்தப்படுகிறது என்பதை அடையாளம் காண்பது தொடர்பான ஒரு திறந்த சிக்கலாகும். இந்தச் சிக்கலுக்கான தீர்வு கணினி தொடர்பான பிற செயல்பாடுகளைப் (இயந்திர மொழிபெயர்ப்பு, கருத்துப்பரிமாற்றம், தேடுபொறிகளை மேம்படுத்துதல், முற்சுட்டுத் தீர்மானம், ஒத்திசைவு, அனுமானம் போன்ற செயல்பாடுகளைப்) பாதிக்கிறது.

கணினி அறிவியல் மற்றும் அது இயக்கும் தகவல் தொழில்நுட்பத்தில், இயற்கை மொழி ஆய்வு மற்றும் இயந்திர கற்றல் ஆகியவற்றைச் செய்வதற்கான கணினிகளின் திறனை வளர்ப்பது நீண்டகால சவாலாக உள்ளது. இயந்திர மொழிபெயர்ப்பில் மிகக் கடினமான பகுதி பொருண்மை மயக்க நீக்கம் ஆகும். பொருண்மை மயக்க நீக்கத்தை அமைப்புப் பொருண்மை மயக்க நீக்கம் சொற்பொருண்மை மயக்க நீக்கம் எனப் பகுத்து உணரலாம்.

ஒரு வாக்கியத்தை நாம் பகுப்பாய்வு செய்யும் போது அவ்வாக்கியத்திற்கு வேறுபட்ட பகுப்பாய்வுகள் சாத்தியமாகும். எடுத்துக்காட்டாக, *மணிக் குமார் தெருவில் ஒரு அழகான பெண்ணைப் பார்த்தான்* என்ற வாக்கியத்தை எடுத்துக்கொண்டால் அதை பின்வருமாறு வேறுபட்ட வகையில் பகுப்பாய்வு செய்யலாம்:

[மணி குமார்]_பெதொ [தெருவில்}_பெதொ [ஒரு பெண்ணைப்]_பெதொ பார்த்தான்.

[மணி]_பெதொ [குமார் தெருவில்]_பெதொ [ஒரு பெண்ணைப்]_பெதொ பார்த்தான்.

[மணி குமார் தெருவில்]_பெதொ [ஒரு பெண்ணைப்]_பெதொ பார்த்தான்.

(பெதொ = பெயர்த்தொடர்) மேற்சொன்ன சாத்தியம் காரணமாக ஒரு வக்கியத்திற்கு ஒன்றுக்கும் கூடுதலான பொருள்கோள்கள் சாத்தியமாகும். இதன் காரணமாக இயந்திர மொழிபெயர்ப்பின் போது ஒரு மூல மொழி வாக்கியத்திற்கு ஒன்றுக்கும் மேற்பட்ட இலக்கு மொழி வாக்கியங்கள் பெறப்படும். இது இயந்திர மொழி பெயர்ப்பின் மிகப் பெரிய சிக்கலாகும்.

அடுத்த சிக்கல் சொற்களின் பல் பொருண்மை அடிப்படையிலானதாகும். எடுத்துக்காட்டாக, *படி* என்ற சொல்லை எடுத்துக் கொண்டால், அதை 'மாடிப் படி', 'எடைப்படி', 'சொன்ன படி', 'புத்தகம் படி' எனப் பலவகையில் பொருள் கொள்ளலாம். பொருண்மை மயக்கதை உணர்ந்து சூழல் அடிப்படையில் சரியாகப் பொருள்கொள்ள உதவும் நமது அறிவைக் கணிப்பொறியாகிய இயந்திரத்திற்குத் தந்தால்தான் தரப்பட்டுள்ள வாக்கியத்திற்குச் சரியான பொருளைப் பெற்று இலக்கு மொழிக்கு அதைச் சரியாக மொழி பெயர்க்கவியலும்.

அமைப்புப் பொருண்மை மயக்கத்தையும் சொற்பொருண்மை மயக்கத்தையும் ஆய்ந்து இயந்திர மொழி பெயர்ப்பில் இதனால் ஏற்படும் சிக்கல்களை நீக்குவதற்கான வழிமுறைகளைக் காண்பதுதான் இவ்வேட்டின் தலையாய நோக்கமாகும். அமைப்புப் பொருண்மை மயக்கநீக்கம் மற்றும் சொற்பொருண்மை மயக்கநீக்கம் பற்றி இதுவரை வெளிவந்த ஆய்வுகளைத் தமிழில் தந்து இது பற்றிய மேலும் பல ஆய்வுகளுக்கு வழிவகுப்பதையும் இவ்வாய்வு நோக்கமாய் கொண்டுள்ளது. இயந்திர மொழிபெயர்ப்புச் சூழலில் அமைப்புப் பொருண்மை மயக்கநீக்கமும் சொற்பொருண்மை மயக்கநீக்கமும் முக்கியத்துவம் வாய்ந்தாக இருந்தாலும் சொற்பொருண்மை மயக்கநீக்கத்தைப் பற்றிதான் அதிகமான ஆய்வுகள் வெளிவந்துள்ளன. மேலும் அமைப்புப் பொருண்மைமயக்கம் தொடரியல் சார் சிக்கலாகத்தான் கருதப்படுகின்றது. இவ்வேட்டில் சொற்பொருண்மை மயக்கநீக்கத்திற்குத்தான் முக்கியத்துவம் தரப்பட்டுள்ளது.

வாக்கிய அமைப்பு நிலையிலும் சொல் நிலையிலும் பொருண்மை மயக்கங்கள் இயந்திர மொழிபெயர்ப்பில் சிக்கல்களை உருவாக்கும். விதி அடிப்படையில் மட்டும் இச்சிக்கல்களைத் தீர்க்க இயலாது. தரவுத்தொகுதிகள் அல்லது பனுவல்களைப் பயன்படுத்தி புள்ளியியல் அடிப்படையிலும் சூழல்களைக் கையாளும் விதிகளைப் பயன்படுத்தியும் இச்சிக்கல்களை ஓரளவுக்குத் தீர்க்கவியலும். பொருண்மை மயக்கச் சிக்கல்களுக்கு முழுவதுமான தீர்வு காண்பது கடினமாகும்.

அமைப்புப் பொருண்மை மயக்கமும் சொற்பொருண்மை மயக்கமும் எவ்வாறு கணினி வழி மொழிபெயர்ப்பில் சிக்கல்களை உருவாக்குகின்றன என்பது ஆயப்பட்டு அச்சிக்கல்களைத் தீர்ப்பதற்கான நெறிமுறைகள் ஆயப்பட்டு நடைமுறைப்படுத்தப்படும். சொற்பொருள் மயக்கநீக்கம் என்பது கணினிசார் வழி சூழலில் சொற்களின் பொருண்மையைக் கண்டுகொள்ளும் ஒரு திறனாகும். சொற்பொருண்மை மயக்கநீக்கம் ஒரு செயற்கை அறிவு-முழுமைச் சிக்கலாகக் (AI-complete problem) கருதப்படுகின்றது; அதாவது, செயற்கை அறிவில் மிகக் கடினமான சிக்கல்களுக்குச் சமமான செயல்பாடாகும். இங்குச் சொற்களின் பொருண்மை மயக்கத்தை நீக்குவதற்கான செயலூக்கங்கள் கருத்தில் கொள்ளப்படும். மேலும் கண்காணிக்கப்பட்ட, கண்காணிக்கப்படாத மற்றும் அறிவு அடிப்படையிலான அணுகுமுறைகள் கருதப்படும். சொற்பொருண்மை மயக்க நீக்க ஒழுங்குமுறைகளின் மதிப்பீடு பல்வேறுபட்ட மயக்க நீக்கச் செயல்படுகளில் பங்கெடுக்கும் ஒழுங்குமுறைகளின் புறவயமான மதிப்பீட்டை நோக்கமாகக்கொண்டு Senseval/Semeval நடவடிக்கைகளின் சூழலில் விவாதிக்கப்படும். இறுதியாயாகப் பயன்பாடுகள், வெளிப்படையான சிக்கல்கள் மற்றும் எதிர்காலத் திசைகள் விவாதிக்கப்படும்.

மனித மொழி பொருண்மை மயக்கமுள்ளது; பல சொற்களை அவை வரும் சூழலைப் பொறுத்துப் பல வழிகளில் பொருள்கோள் செய்யவியலும். எடுத்துக்காட்டாக, பின்வரும் வாக்கியங்களை எடுத்துகொள்ளவும்:

அ. நான் அவன் சொன்னதைக் கேட்டேன்.

ஆ. நான் அவன் சொன்னதைக் காதால் கேட்டேன்.

இ. நான் அவன் சொன்னதைக் கேட்டு அதன் படி நடந்தேன்.

முதல் வாக்கியத்தில் *கேட்டேன்* என்பது பொருண்மை மயக்கம் உடையதாய் இருக்கின்றது. *கேள்* என்ற வினை முதல் வாக்கியத்தில் 'hear' என்ற பொருண்மையையும் 'oblige' என்ற பொருண்ண்மையும் தந்து மயக்கம் உடையதாய் இருக்கின்றது. இரண்டாவது வாக்கியத்தில் *கேள்* என்ற வினை 'hear' என்ற பொருண்மையை மட்டும் குறித்து நிற்கின்றது. மூன்றவது வாக்கியத்தில் *கேள்* என்ற வினை 'oblige' என்ற பொருண்மையை மட்டும் குறித்து நிற்கின்றது.

துரதிருஷ்டவசமாக, சூழலில் சொல் பெறும் குறிப்பிட்ட பொருண்மையின் கண்டுபிடிப்பு மேலீடாகத்தான் எளிதாகத் தோன்றுகின்றது. பொரும்பாலன செயல்பாடுகளில் மனிதர்கள் மொழியிலுள்ள மயக்கங்களைப் பற்றிச் சிந்திப்பதே இல்லை; அதே சமயம் இயந்திரங்கள் (கணினிகள்) அமைப்பாக்கம் செய்யப்படாத பனுவல் தகவல்களைப் பகுப்பாய்வு செய்யவேண்டும்; மற்றும் அவற்றை உள்ளுறும் பொருண்மையை நிர்ணயிப்பதற்காகப் பகுத்தாயப்படவேண்டிய தரவு அமைப்பாக மாற்ற வேண்டும். சூழலில் சொற்களின் பொருண்மைகளின் கணினிசார் கண்டுபிடிப்பு சொல் மயக்கநீக்கம் எனப்படும். இத்தகைய செயல்பாடுகளை விதி அடிப்படையிலும் புள்ளியியல் அடிப்படையிலும் செயல்படுத்தலாம்.

நாம் இயந்திர மொழிபெயர்ப்பின் சிக்கலைக் கூறும் போது சொற்பொருண்மை மயக்கநீக்கத்தின் முக்கியத்துவம் தெளிவாகும்: *saw* என்ற ஆங்கிலச் சொல் சூழல் அடிப்படையில் தமிழில் 'பார்' அல்லது 'அரம்' என்று மொழிபெயர்க்கப்படலாம். இம்மாதிரியான நேர்வுகள் ஆயிரக்கணக்கில் உள்ளன; இது சொற்பொருள் மயக்கநீக்கத்தின் வரலாற்றுப் பயன்பாடாகும். பனுவலின் தானியக்க மொழிபெயர்ப்பில் (automated text translation) பொருண்மைமயக்க நீக்கம் முக்கியமான பங்களிப்பு செய்யும்.

இயற்கை மொழி ஆய்வுகளில் தரவுத்தொகுதிகள் அல்லது பனுவல்களில் வரும் பொருண்மை மயக்கங்களை, குறிப்பாகச் சொற்பொருண்மை மயக்கநீக்கத்திற்கான வழிமுறைகள் பல பன்பாட்டில் உள்ளன. தமிழில் சொற்பொருண்மை மயக்கநீக்கத்திற்கான வழிமுறைகள் அதிகமாக மேற்கொள்ளப்படவில்லை. எடுத்துக்காட்டுகள் என்ற நிலையில் தான் இம்முயற்சிகள் அமைகின்றன. இவ்வாய்வேட்டில் தமிழில் சொற்பொருள் மயக்கநீக்கத்திற்கான வழிமுறைகள் ஆயப்பட்டு விளக்கப்பட்டுள்ளன.

பொருண்மை மயக்கநீக்கம் இயற்கை மொழி ஆய்வில் பெரும் பங்கு வகிக்கின்றது. இயந்திர மொழிபெயர்பு ஒழுங்கு முறை, தகவல் மீட்பு ஒழுங்குமுறை, தகவல் தேடல்

ஒழுங்குமுறை, கேள்வி-விடை ஒழுங்குமுறை போன்றவற்றில் சொற்பொருண்மை மயக்கநீக்கம் இன்றியமையாத பங்களிப்பு செய்கின்றது. தமிழுக்கு இத்தகைய ஆய்வு ஒரு வரப்பிரசாதமாக அமையும்.

இயல் 2

அமைப்புப் பொருன்மை மயக்கமும் சொற்பொருண்மை மயக்கமும்

2.0 முன்னுரை

மயக்கம் (ambiguity) அல்லது தெளிவின்மை என்பது ஒரு வகை அர்த்தம். இதில் ஒரு சொற்றொடர், அறிக்கை அல்லது தீர்மானம் வெளிப்படையாக வரையறுக்கப்படவில்லை. இது பல விளக்கங்களை நம்பத்தகுந்ததாக ஆக்குகிறது. மயக்கம் அல்லது தெளிவின்மையின் பொதுவான அம்சம் நிச்சயமற்ற தன்மை. இது எந்தவொரு யோசனை அல்லது அறிக்கையின் ஒரு பண்பு ஆகும். அதன் விதிமுறை அல்லது செயல்முறைக்கு ஏற்ப வரையறுக்கப்பட்ட எண்ணிக்கையிலான வழிமுறைகளால் திட்டவட்டமாக தீர்க்க முடியாது. (ambiguity என்ற வார்த்தையின் ambi பகுதி "இரண்டு அர்த்தங்கள்" என்பது போன்று "இரண்டு" என்ற கருத்தை பிரதிபலிக்கிறது.)

மயக்கம் அல்லது தெளிவின்மை என்ற கருத்து பொதுவாக தெளிவற்ற தன்மைக்கு (vagueness) முரணானது. தெளிவற்ற நிலையில், குறிப்பிட்ட மற்றும் தனித்துவமான விளக்கங்கள் அனுமதிக்கப்படுகின்றன (சில உடனடியாகத் தெரியாவிட்டாலும்), அதேசமயம் தெளிவற்ற தகவல்களுடன், விரும்பிய அளவிலான விவரக்குறிப்பில் எந்த விளக்கத்தையும் உருவாக்குவது கடினம்.

மயக்கம் சொல்நிலையிலோ தொடரியல் நிலையிலோ ஏற்படும். சொல்நிலையில் ஏற்படும் மயக்கம் சொற்பொருண்மை மயக்கம் என்றும் தொடரியல் நிலையில் ஏற்படும் மயக்கம் அமைப்புப் பொருண்மை மயக்கம் என்றும் அழைக்கப்படும். இவ்வியலில் அமைப்புப் பொருண்மை மயக்கம் பற்றியும் சொற்பொருண்மை மயக்கம் பற்றியும் விரிவான விளக்கம் தரப்பட்டுள்ளது. இவ்விரு மயக்கங்களின் வகைகள் எடுத்துக்காட்டுகளுடன் ஆயப்பட்டுள்ளது. தமிழ்த் தரவுதொகுதிகளில் இவ்விரு மயக்கங்கள் வந்து பொருள்கோண்மையை எவ்வாறு பாதிக்கின்றது என்பது பற்றியும் விரிவாக ஆயப்பட்டுள்ளது.

2.1 பொருண்மை மயக்கம்

ஒரு சொல்லோ தொடரோ வாக்கியமோ ஒன்றிற்கும் மேற்பட்ட பொருண்மைகளை உணர்த்துவதாய் இருந்தால் அவை பொருண்மை மயக்கம் உள்ளனவாகும் எனலாம். எடுத்துகாட்டாக, *கால்* என்ற சொல் ஒன்றின் நான்கில் ஒரு பகுதியையும் உடல் உறுப்பான

காலையும் குறிப்பிடும். இவ்வாறு சொற்கள் பொருண்மை மயக்கம் தருவது சொற்பொருண்மை மயக்கம் எனப்படும். சொற்பொருண்மை மயக்கம் தொடரிலும் வாக்கியத்திலும் பொருண்மை மயக்கத்தை உருவாக்கும்.

கால் பகுதி .'உடலின் ஒரு உறுப்புப் பகுதி, நான்கில் ஒரு பகுதி'

1. அவன் சாப்பிட்டது கால் பகுதி

இருப்பினும் தொடரும் வாக்கியமும் சொற்பொருண்மை மயக்கம் உள்ள சொல் இல்லாவிடினும் பொருள் மயக்கம் தரலாம். எடுத்துக்காட்டாகப் பின்வரும் தொடர்களை எடுத்துக்கொள்ளவும்.

2. வெள்ளை மருந்துக் குப்பி

3. பழைய மாணவர் விடுதி

முதல் வாக்கியம் 'மருந்துள்ள வெள்ளை நிறக் குப்பி' என்றும் 'வெள்ளை மருந்துள்ள குப்பி' என்றும் பொருள் மயக்கம் காட்டும். இதுபோன்று இரண்டாவது வாக்கியமும் 'பழைய மாணவர்கள் வசிக்கும் விடுதி' என்றும் 'மாணவர்கள் வசிக்கும் பழைய விடுதி' என்றும் பொருள் மயக்கம் தரும்.

பொருண்மை மயக்கம் என்ற கருத்துச் சாயலுக்குத் தத்துவப் பயன்பாடுகள் உண்டு. எடுத்துக்காட்டாக, பொருண்மை மயக்கத்தை அடையாளம் காண்பது தத்துவச் சிக்கல்களைத் தீர்க்கும். ஒருவர் எவ்வாறு இருவர் *ஒற்றைக்கண்ணன்* என்பதற்கு ஒரே கருத்தைக் கொண்டுள்ளனர் என்று ஆச்சரியப்படக்கூடும். ஒருவர் 'கருத்து' என்பதை ஒரு குறிப்பிட்ட உளவியல் நேர்வு அதாவது ஒரு மன உருப்படுத்தம் என்ற அர்த்தத்தை அருவத்தன்மையான பங்கிடவியலும் கருத்துரு என்ற அர்த்தத்திலிருந்து வேறுபடுத்துவதுவரை அது புதிராக இருக்கும். மாறாக, தன்விருப்பார்ந்த/உள்நோக்கமற்ற பொருண்மை மயக்கக் கூற்றுகள் எளிய தீர்வுகளுக்கு வழிவகுக்கும். பதிலின் பகுதி பொருண்மை மயக்கம் என்ற கருத்துச்சாயலைத் தெளிவற்றது, நிச்சயமற்றது, குழப்பம் தருவது என்பவைகளுடன் அடையாளம் காண்பதுதான்.

2.2 பொருண்மை மயக்கத்தின் வகைகள்

மக்கள் பொருண்மை மயக்கம் காட்டுகின்றார்கள் என்று கூறுவது அவர்கள் எவ்வாறு மொழியைப் பயன்படுத்துகின்றனர் என்பதைக் குறிக்கும் என்றாலும், பொருண்மை மயக்கம் என்பது மொழி வெளிப்பாடுகளின் பண்பாகும். ஒரு சொல்லோ தொடரோ வாக்கியமோ ஒன்றிற்கும் மேற்பட்ட பொருண்மைகளைக் கொண்டிருந்தால் அவை பொருண்மை மயக்கம் கொண்டுள்ளன என்று பொருள்படும். வெளிப்படையாக இந்த வரையறை விளக்கம்

பொருண்மைகள் என்றால் என்ன அல்லது ஒரு வெளிப்பாடு ஒன்றொ அல்லது ஒன்றிற்கும் மேற்பட்ட பொருண்மைகளைக் கொண்டிருக்கும் என்பது என்ன என்று கூறவில்லை. ஒரு குறிப்பிட்ட மொழிக்கு இந்தத் தகவல் இலக்கணத்தால் தரப்படும்; இது முறையாக வடிவங்களைப் பொருண்மைகளுடன் பொருத்தும் மற்றும் பொருண்மை மயக்கம் உள்ள வடிவங்களை ஒன்றிற்கும் மேற்பட்ட பொருண்மைகளுடன் பொருத்தும்.

இருவகையான பொருண்மை மயக்கங்கள் முக்கியமாகப் பேசப்படுகின்றன. அவையாவன சொற்பொருண்மை மயக்கம் மற்றும் அமைப்புப் பொருண்மை மயக்கம்.

2.2.1 சொற்பொருண்மை மயக்கம்

சொற்பொருண்மை மயக்கம் மிகப் பொதுவானதாகும். எடுத்துக்காட்டுகள் *படி, குடி, மடம்* என்ற பெயர்களையும் *பிடி, அறை, முடி, சுடு* என்ற வினைகளையும் *நீண்ட, விரிந்த, குறைந்த, வெளுத்த, கறுத்த* என்ற பெயரடைகளையும் உள்ளடக்கும். பொருண்மை மயக்கத்திற்கு பல சோதனைகள் உள்ளன. ஒரு சோதனை *கடினமான* என்பதற்கு *மெதுவான, எளிதான* என்ற இரண்டும் எதிர்சொற்கள் உள்ளன. பின்வரும் வாக்கியத்தைக் கருத்தில் கொள்ளவும்.

4.அ.*கடினமான மிட்டாயைக் கடிக்க முடியாது.*

4.ஆ.*கடினமான சொல்லுக்குப் பொருள் கூற இயலாது.*

முன்னர் தந்த பொருள் மயக்கத்திற்கான எடுத்துக்காட்டு ஒரு சொல்லுக்கு ஒன்றுக்கும் மேற்பட்ட பொருண்மை காரணமாக விளைந்தது. இருப்பினும் எப்போது நமக்கு ஒரு சொல் மட்டும் இருக்கும் என்பதில் தெளிவில்லை. *படி* என்ற வினையும் *படி* என்ற பெயரும் ஒரே உச்சரிப்பைக் கொண்டிருந்தாலும் இரண்டு சொற்கள் ஆகும். இவை ஒப்புரு மொழிக்கு எடுத்துக்காட்டாகும். *கடி* என்ற வினையும் *கடி* என்ற பெயரும் ஒப்புரு மொழிகளா? *முதல் மாணவனில்* வரும் *முதல்* என்பதும் *ஐந்து முதல்* என்பதில் வரும் *முதல்* என்பதும் ஒன்றா? என்ற கேள்விகள் எப்பொழுதும் எழுப்பப்படும். ஒன்று பொருள் மயக்கம் காட்டுவது மற்றொன்று ஒப்புரு மொழிகள் என்று கூறுவதில் ஒருமித்த கருத்து இல்லை. ஒரு வேளை இந்த வேறுபாடு தன்னிச்சையானதாகும்.

சில வேளைகளில் ஒரு சொல்லின் பொருள் பிற சொல்லிலிருந்து ஆக்கப்பட்டிருக்கும். எடுத்துக்காட்டாக *காண்* என்ற வினைச் சொலின் பொருள் *கண்* என்ற சொல்லின் பொருளிலிருந்து உருவாகி இருக்கக்கூடும். இதே போன்று *எரி, உடை, ஒடி* என்ற சொற்களின் செயப்படுபொருள் குன்றாவினைப் பொருண்மைகள் செயப்படுபொருள் குன்றிய வினைப்பொருண்மைகளிலிருந்து ஆக்கப்பட்டிருக்கலாம். இப்போது இந்த நேர்வுகள்

ஒவ்வொன்றிலும் ஆக்கப்பட்ட பொருண்மை அச்சொல்லின் இரண்டாவது பொருண்மையாகத் தகுதி பொறாது; ஆனால் இது ஆக்கப்படாத பொருண்மையின் மீதான சொல்லியல்சார் செயற்பாடாகும். இந்த நிகழ்வு குறிப்பிட்ட சொற்களுக்குச் சிறப்பானது என்றில்லாமல் முறையானது மற்றும் பொதுவானது என்ற அளவில் இந்த வாதம் சாத்தியமானதாகும். சொற்பொருண்மை இயலுக்கு இம்மாதிரியான பொருண்மை நிகழ்வுகளை அடையாளம் காணும் மற்றும் பண்பாக்கம் செய்யும் பொறுப்பு உள்ளது. மேலும் *செய், போடு* என்ற வினைகளும் *மேல், கீழ்* என்ற பின்னுருபுகளும் காட்டும் பொதுவான மற்றும் மிக நெகிழ்வான பொருண்மை நடத்தையை விளக்கும் பொறுப்பும் உள்ளது.

2.2.1.1 சொற்பொருண்மை மயக்கத்தின் வகைகள்

சொற்பொருண்மை மயக்கங்கள் மூன்று அடிப்படை வகைகளில் வேறு படும்: சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம் (parts-of-speech or category ambiguity), ஒப்புருமொழியம்சார் (சொற்போலி) பொருண்மை மயக்கம் (ambiguity due to homography), பல்பொருள் ஒருமொழியம்சார் (பல்பொருண்மை) பொருண்மை மயக்கம் (ambiguity due to polysemy), மாற்றப் பொருண்மை மயக்கம் (ambiguity due to transfer or transfer ambiguity).

2.2.1.1.1 சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம்

சொற்பொருண்மை மயக்கத்தின் நேரடியான வகை சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம் ஆகும்: ஒரு சொல் ஒன்றிற்கும் மேற்பட்ட இலக்கண அல்லது தொடரியல் வகைப்பாட்டிற்கு ஒதுக்கப்படும். தமிழில் இத்தகைய பல எடுத்துகாட்டுகள் உள்ளன. *பச்சை* என்பது பெயராகவும் பெயரடையாகவும் வரும்; *சுடு* என்பது வினையாகவும் பெயரடையாகவும் வரும்; *கடி* என்பது வினையாகவும் பெயராகவும் வரும். *மேலே, கீழே* என்ற சொற்கள் பெயராகவும் வினையடைகளாகவும் பின்னுருபுகளாகவும் வரும்.

5.அ. அவன் மேலே இருக்கிறான். (பெயர்)

5.ஆ. அவன் மேலே சென்றான். (வினையடை)

5.இ. அவன் மேசை மேலே நிற்கின்றான். (பின்னுருபு)

2.2.1.1.2 ஒப்புருமொழியம் சார், பல்பொருள் ஒருமொழியம் சார் பொருண்மை மயக்கம்

ஒரு சொல்லுக்கு இரண்டோ அதற்கு மேலோ பொருண்மைகள் இருந்தால் இரண்டாவது வகைச் சொற்பொருண்மை மயக்கம் ஏற்படும். இரண்டோ அதற்கு மேற்பட்ட சொற்கள் முற்றிலும்

வேறுபட்ட பொருண்மையைக் கொண்டிருப்பது ஒப்புருமொழியம் (homography) ஆகும். எடுத்துகாட்டாகப் *படி* என்பது மாடிப்படியையும் கற்பதையும் குறிப்பிடும்.

6. அவன் தந்தை அவனிடம் நீ நன்றாகப் படி என்று கூறினார்.

7. அவன் படி வழியாக மடியில் ஏறினான்.

பின்வரும் வாக்கியம் ஒப்புருமொழியம் காரணமாகப் பொருண்மை மயக்கம் காட்டும்.

8. அவன் காட்டியதும் படி.

மேற்கண்ட வாக்கியம் 'what he showed was also step' மற்றும் 'you read also what he showed' என இரு அர்ததங்களை வெளிப்படுத்திப் பொருண்மை மயக்கம் காட்டும்.

அட்டை என்பது புத்தக அட்டையையும் அட்டைப் பூச்சியையும் குறிப்பிடும்.

9. அவன் புத்தகத்தின் அட்டையைக் கிழித்து எறிந்தான்.

10. அவன் அந்த அட்டையைக் கொன்றான்.

பின்வரும் வாக்கியம் ஒப்புருமொழியம் காரணமாகப் பொருண்மை மயக்கம் காட்டும்.

11.அவள் அட்டையைப் பார்த்தாள்

மேற்கண்ட வாக்கியம் 'she saw the wraper' மற்றும் 'she saw the leech' என்ற இரு அர்த்தங்களை வெளிப்படுத்திப் பொருண்மை மயக்கம் காட்டும்.

பல்பொருள் ஒருமொழிகள் (polysememes) தொடர்புள்ள ஒரு தொடர்ச்சியான பொருண்மைகளை வெளிப்படுத்தும். சொற்கள் தமது உருவக நீட்சியாலும் மாற்றத்தாலும் புதிய பொருண்மைகளை வெளிப்படுத்தும். எடுத்துக்காட்டாக, *கிளை* என்பது மரக்கிளையையும் வங்கிக் கிளையையும் குறிப்பிடும்; *நட* எனபது நடக்கும் இயக்கத்தையும் நிறுவனத்தின் செயல்பாட்டையும் குறிப்பிடும்.

12.அ.காகம் அந்த மரத்தின் கிளையில் உட்கார்ந்திருக்கின்றது.

12.ஆ. அவன் அந்த வங்கியின் கிளையில் வேலை செய்கிறான்.

13.அ. தினமும் காலையில் பள்ளிக்கு நடந்து செல்கின்றான்.

13.ஆ. அந்த நிறுவனம் நன்றாக நடந்து கொண்டிருக்கின்றது.

பின்வரும் வாக்கியங்கள் பல்பொருள் ஒருமொழியம் காரணமாக பொருண்மை மயக்கம் காட்டும்.

13.இ.இங்கு புள்ளிமான் நடந்து கொண்டிருக்கிறது.

இவ்வாக்கியம் 'இந்த தியேட்டரில் புள்ளிமான் என்ற சினிமா நடந்துகொண்டிருக்கின்றது' எனவும் பொருள் தந்து பொருண்மை மயக்கம் காட்டும்.

ஓடு என்பது மனித இயக்கத்தையும் ஆறு ஒழுகுவதையும் குறிப்பிடும்.

14.அ. அவன் விரைவாக ஓடுகிறான்.

14.ஆ. காவிரி ஆறு தஞ்சாவூர் வழியாக ஓடுகின்றது.

ஆனால் பின்வரும் வாக்கியம் பல்பொருள் ஒருமொழியம் காரணமாகப் பொருண்மை மயக்கம் காட்டும்.

14.இ.அது வேகமாக ஓடுகிறது

கண் என்பது விலங்கினங்களின் கண்ணையும் தேங்காய் கண்ணையும் குறிப்பிடும்.

15.அ. அவன் தன் கண்களை மூடினான்.

15.ஆ. தேங்காய்க்கு மூன்று கண்கள் உண்டு.

பின்வரும் வாக்கியதில் *கேள்* என்ற சொல் செவிப்புலத்தையும் வினவுதலையும் குறித்து நின்று சொற்பொருண்மை மயக்கம் காட்டும்.

16. ராதை ராஜா கேட்டதை அவனிடம் கூறினாள்.

இது சொற்பொருண்மை அடிப்படையில் பின்வருமாறு இரு அர்த்தங்களைத் தந்து சொற்பொருண்மை மயக்கம் காட்டும்.

17. ராதை ராஜா தன் காதால் கேட்டதை அவனிடம் கூறினாள்.

18. ராதை ராஜா வினவியதை அவனிடம் கூறினாள்.

சில வேளைகளில் ஒப்புருமொழி இணைகளில் ஒன்றின் பயன்பாடு மற்றதைவிட கூடுதலாக இருக்கக்கூடும். இந்நேர்வில் ஒப்புருமொழிகளின் பொருண்மை மயக்கத்தை பனுவல் வகை அடிப்படையில் நீக்க இயலும்; எனவே வழக்கமில்லாத பயன்பாடு மொழிபெயர்க்க வேண்டிய பனுவலின் விசயத்திற்குப் பொருந்தினால் ஒழிய அகராதியிலிருந்து விலக்கி வைக்கப்படும்.

இயந்திர மொழிபெயர்ப்பில் ஒப்புருமொழியத்தையும் (homography) பல்பொருள் ஒருமொழியத்தையும் (polysemy) பெரும்பாலும் ஒன்றுப்போல் கையாள இயலும். ஏனென்றால் இங்கு நோக்கம் எழுத்தப்பட்ட சொல்லுக்குச் சூழல் அடிப்படையில் பொருண்மையைக் கண்டுபிடிப்பது ஆகும். மேற் சொன்ன வழியில் வேறுபட்ட சொல்வகைப்பாட்டைச் சார்ந்த ஒப்புரு மொழிகளின் பொருண்மை மயக்கத்தைத் நீக்க இயலும்; ஆனால் ஒப்புரு மொழிகள் ஒரே சொல்வகைப்பாட்டைச் சார்ந்திருந்தால் தொடரியல் பகுப்பாய்வு மட்டும் போதுமானதல்ல.

பொதுவான ஒரு அணுகுமுறை 'மனிதவினம்', 'பெண்ணினம்', 'திரவம்' போன்ற பொருண்மைப் பண்புக்கூறுகளைத் தந்து தேர்வுக்கட்டுப்பாடு அடிப்படையில் எந்தப் பண்புக்கூறு தரப்பட்ட தொடரியல் கட்டுமானத்துடன் இணங்கும் என்று குறிப்பிடுவதாகும். எடுத்துக்காட்டாக, *குடி* என்ற வினை ஒரு 'விலங்கின' எழுவாயைக் கொண்டிருக்கவேண்டும்.

தொடர்ந்து பயன்படுத்த இயலும் பொருண்மைப் பண்புகூறுகளின் குழுமத்தை உருவாக்குவதுதிலும் இப்பண்புக்கூறுகள் அடிப்படையில் பெயர்கள் மற்றும் வினைகள் இவற்றின் தேர்வுக் கட்டுப்பாடுகளைக் குறிப்பிடுவதிலும் சிக்கல்கள் உள்ளன. இருப்பினும் அவைகள் மொழிபெயர்ப்பு ஒழுங்கு முறைகளில் பெரும்பாலும் வேற்றுமைப் பங்கெடுப்பாளர்கள் (case roles) தகவலுடன் சேர்த்துப் பரவலாகப் பயன்படுத்தப்படுகின்றன. எடுத்துக்காட்டாக, *அட்டை* என்ற ஒப்புரு மொழியை எடுத்துக்கொள்ளவும்; இதற்கு முன்னர் கூறியபடி '(புத்தக) அட்டை' மற்றும் '(மர) அட்டை' என்ற இரு பொருண்மைகள் உண்டு. இந்த இரண்டு பொருண்மைகளையும் பின்வரும் வாக்கியங்களில் உள்ள பொருத்தமான சேர்ந்துவருகைக் கட்டுப்பாட்டுப் பண்புக்கூறுகளை (selection restriction features) விளக்கி வேறுபடுத்தலாம்.

19. புத்தகத்தின் அட்டை கிழிந்து விட்டது.

20. அந்த அட்டை தரையில் ஊர்ந்து கொண்டிருந்தது.

கிழி என்ற வினை *அட்டை* போன்ற கிழியக்கூடிய பொருளை எழுவாயாக ஏற்கும். *ஊர்* என்ற வினை ஊர்ந்து செல்லக்கூடிய அட்டைப் பூச்சியை எழுவாயாக ஏற்கும்.

ஒப்புருமொழியம் இலக்கண ஆய்வு அடிப்படையிலான வேறுபாடு அடிப்படையிலும் வெளிப்படும் (இலக்கணப்போலி). பின்வரும் எடுத்துக்காட்டுகள் இதை உணர்த்தும்.

21. அவன் கடலை தின்று மகிழ்ந்தான்

22. அவன் கடலைக் கண்டு மகிழ்ந்தான்.

21-ஆவது வாக்கியத்தில் *கடலை* என்பது *கடலை* என்ற தின்பண்டத்தைக் குறிப்பிடும் தனிச்சொல்லாகும். 9-ஆவது வாக்கியத்தில் *கடலை* என்பது *கடல்+ஐ* எனப் பகுத்தாய இயலும் சமுத்திரத்தைக் குறிப்பிடும் திரிபுற்ற சொல்லாகும். இத்தகைய ஒப்புருமொழியமும் பொருண்மை மயக்கத்தை விளைவிக்கும்.

23. அவன் பார்த்தது கடலை.

மேற்கண்ட வாக்கியத்தி கடலை என்பது சமுத்திரத்தையும் உண்ணப்படும் கடலையும் ஒரே சமயத்தில் குறித்துப் பொருண்மை மயக்கம் காட்டும்.

பின்வரும் வாக்கியங்களில் இரு சொற்களின் வேறுபட்ட இலக்கண வடிவுகள் ஒன்று போல் அமைந்து ஒப்புருமொழியம் காட்டும்.

23. அவன் துணி நெய்தான்.

24. அவன் நெய்தான் விருப்புகிறான்

23-ஆவது வாக்கியத்திலுள்ள *நெய்தான்* என்பதும் 24-ஆம் வாக்கியத்திலுள்ள *நெய்தான்* என்பதும் இருவேறு சொற்களின் இருவேறு இலக்கண வடிவங்கள். முதல் வாக்கியம் 'he weaved (a cloth)' என்றும் இரண்டாவது வாக்கியம் 'he likes only ghee' என்றும் பொருள்பட்டு வேறுபடும். ஒப்புரு மொழியம் காரணமாகப் பொருண்மை மயக்கம் விளையும்.

25. அவள் கேட்டதும் நெய்தான்.

25-ஆம் வாக்கியம் 'what she asked was also ghee' எனவும் 'he weaved as soon as she asked' பொருண்மை மயக்கம் காட்டும்.

வரலாற்று மொழியியல் அடிப்படையில் சொற்களின் திரிபு வடிவங்கள் வேறுபட்ட இலக்கணச் செயல்பாடுகளைப் பெற்று வேறுபட்ட சொல்வகைப்பாடுகளைக் குறித்துநிற்கும் அதாவது வேறுபட்ட பொருண்மைகளைக் குறித்துநிற்கும். இருப்பினும் அவை ஒரே வடிவில் இருப்பதால் அவற்றை ஒப்புருமொழிகளின் வகையாகக் கருத இயலும். குறிப்பாக வினைகளின் திரிபுற்ற வடிவங்களிலிருந்து உருவாக்கப்பட்ட அல்லது மாற்றுச் செயல்பாடு பெற்ற முன்னுருபுகள் (post postions) இத்தகைய பொருண்மை மயக்கம் காட்டும். எடுத்துக்காட்டாக *இருந்து, பற்றி, குறித்து, ஒட்டி, கொண்டு, வைத்து, சுற்றி, நோக்கி, முந்தி, விட, கூட* என்ற முன்னுருபுகள் அவற்றின் மூல வினைகளின் திரிபுற்ற ஆனால் வகைப்பாடு (செயல்பாடு) மாறாத வடிவங்களுடன் பொருண்மை மயக்கம் காட்டும்.

25அ. அவன் வீட்டில் இருந்து வெளியேறினான். (முன்னுருபு)

25ஆ. அவன் வீட்டில் இருந்து வந்தான் (வினைத்திரிபு வடிவம்)

26அ. அவன் அவளைப் பற்றி பேசினான். (முன்னுருபு)

26ஆ. அவன் அவள் கையைப் பற்றி முத்தமிட்டான். (வினைத்திரிபு வடிவம்)

27அ. அவன் அவளைக் குறித்து பேசினான். (முன்னுருபு)

27ஆ. அவன் அவள் சொல்வதைக் குறித்து வந்தான். (வினைத்திரிபு வடிவம்)

28அ. அவன் அந்த தலைப்பை ஒட்டி பேசினான். (முன்னுருபு)

28ஆ. அவன் போஸ்டர் ஒட்டி பிழைக்கின்றான். (வினைத்திரிபு வடிவம்)

29அ. அவன் கத்தி கொண்டு அதை வெட்டினான். (முன்னுருபு)

29ஆ. அவன் கத்தியால் வெட்டிக் கொண்டு பேசினான். (வினைத்திரிபு வடிவம்)

30அ. அவன் கத்தி வைத்துப் பழம் வெட்டினான். (முன்னுருபு)

30ஆ. அவன் பணம் வைத்துக் கொண்டு சூதாடினான். (வினைத்திரிபு வடிவம்)

31.அ அவன் வீட்டைச் சுற்றி மரங்கள் நிற்கின்றன. (முன்னுருபு)

31ஆ. அவன் அவளையே சுற்றி வருகின்றான். (வினைத்திரிபு வடிவம்)

32அ. அவன் அவளை நோக்கி நடந்தான். (முன்னுருபு)

32ஆ. அவள் முகத்தை நோக்கிச் சிரித்தான். (வினைத்திரிபு வடிவம்)

33அ. அவன் அவளுக்கு முந்தி அங்கு வந்தான். (முன்னுருபு)

33ஆ. அவன் அவளை முந்தி நடந்துகொண்டிருந்தான். (வினைத்திரிபு வடிவம்)

34அ. அவன் அவளை விட நல்லவன். (முன்னுருபு)

34ஆ. அவன் அவளை விட விருப்பவில்லை. (வினைத்திரிபு வடிவம்)

35அ. அவன் அவள் கூட வந்தான். (முன்னுருபு)

35ஆ. அவனது மகிழ்ச்சி கூட ஆர்ப்பரித்தான்.

என் என்ற வினையின் திரிபுற்ற *என்று* என்ற வடிவம் மூன்று வேறுபட்ட இலக்கணச் செயல்பாடுகளில் வந்து (இலக்கணப்போலி) ஒப்புருமொழியம் காட்டும்.

36அ. அவள் நல்லவள் என்று நினைத்தேன். (*என்று* நிரப்பானாக வருகின்றது)

36ஆ. அவள் திடீர் என்று வந்தாள் (*என்று* வினையடையாக்கியாக வருகின்றது)

36இ. அவள் என்று வருகிறாள். (*என்று* காலத்தைக் குறிந்து வருகின்றது)

வரும் என்ற வடிவம் எதிர்கால வினை முற்றாகவும் பெயரெச்சாமாகவும் செயல்படும்.

37அ. அது நாளை வரும்.

37ஆ. அது வரும் நாள் எனக்குத் தெரியாது.

-*அது* ஒட்டப்பட்ட வினைவடிவம் மூன்று வேறுபட்ட இலக்கணச் செயல்பாடுகளில் வந்து (இலக்கணப் போலி) ஒப்புருமொழியம் காட்டும்.

38அ. அது நேற்று வந்தது (*வந்தது* வினைமுற்று வடிவம்)

38ஆ. அவன் வந்தது எனக்குத் தெரியாது (*வந்தது* வினையாலணையும் பெயர்)

38இ. அந்த செய்தித்தாள் நேற்று வந்தது (வந்தது அது மாற்றுப் பெயராக வினை)

தேர்வுக் கட்டுப்பாடு, சூழல் கட்டுப்பாடு, சூழல் அமைப்பு, சேர்ந்துவருகை என்பன இத்தகைய ஒப்புரு மொழிகளின் பொருண்மை வேறுபடுத்தும்.

2.2.1.1.3 மாற்றப் பொருண்மை மயக்கம்

ஒப்புருமொழியம், பல்பொருள் ஒருமொழியம் என்பன சொற்பொருண்மை மயக்கங்களுக்கு வழிவகுக்கும். அவை முதன்மையாக மூல மொழிப் பனுவல்களின் பகுப்பாய்வில் சிக்கலை விளைவிக்கின்றன. இயந்திர மொழிபெயர்ப்பில் மாற்றப் பொருண்மை மயக்கங்கள் (மொழிபெயர்ப்பு பொருண்மை மயக்கங்கள்) உள்ளன. ஒரு மூல மொழிச் சொல்லை இலக்கு மொழியில் பல சொற்களாலோ வெளிப்பாடுகளாலோ மொழிபெயர்க்க இயலும். மூல மொழிச் சொல் பொருண்மை மயக்கம் காட்டாது அல்லது பொருண்மை மயக்கம் இயல்மொழி பேசுபவர்களால் உணர்ந்து கொள்ளப்படாது இருக்கலாம். இச்சொல் பிற மொழி பேசுபவரின் புரிந்துகொள்ளலின் அடிப்படையில் தான் பொருண்மைமயக்கம் காட்டுகின்றது. எனவே இது மொழிபெயர்ப்பின் சிக்கலாகும். எடுத்துக்காட்டாக *listen* என்ற ஆங்கிலச் சொல் தமிழில் *கேள்* என்று மொழிபெயர்க்கப்படும் போது அதன் ஒப்புருமொழியத் தன்மை பொருண்மை மயக்கத்திற்கு வழி வகுக்கும்.

2.2.2 அமைப்புப் பொருண்மை மயக்கம்

ஒரு தொடருக்கோ அல்லது வாக்கியத்திற்கோ ஒன்றிற்கும் மேற்பட்ட அக அமைப்புகள் இருந்தால் அமைப்புப் பொருண்மை மயக்கம் ஏற்படும். எடுத்துக்காட்டாக *இந்திய வரலாற்று ஆசிரியர்* என்ற தொடர் [*இந்திய வரலாற்று*] *ஆசிரியர்* என்றும் *இந்திய* [*வரலாற்று ஆசிரியர்*] என்றும் பகுத்தாயப்பட்டு பொருண்மை மயக்கம் காட்டும். இது போன்று வயதான ஆண்களும் பெண்களும் என்பது [*வயதான*] [*ஆண்களும் பெண்களும்*] என்றும் [*வயதான ஆண்களும்*] [*பெண்களும்*] என்றும் பகுத்தாயப்பட்டு பொருண்மை மயக்கம் காட்டும். இவ்வகையிலான பொருண்மை மயக்கங்கள் அக அமைப்பு நிலைக்கு வலுவான சான்றைத் தருகின்றன. *கோழி சாப்பிடுவதற்குத் தயார்* என்பது 'கோழி தனது உணவை உண்பதற்குத் தாயார்' எனவும் 'கோழி அதைப் பிறர் உண்பதற்குத் தயாரான நிலையில் உள்ளது' எனவும் பொருண்மை மயக்கம் காட்டும். இத்தகைய பொருண்மை மயக்கங்களை விளக்க மேற்கண்ட புற அமைப்பு வாக்கியத்திற்கு குறைந்தது இரண்டு அக அமைப்புகள் தரவேண்டி வரும்.

எப்போது அமைப்புப் பொருண்மை மயக்கம் வரும் என்பது எப்பொழுதும் தெளிவாக இருப்பதில்லை. பின்வரும் வாக்கியத்தை கருத்தில் கொள்ளவும்.

ராஜாவுக்கு இராமனை விட பணக்காரனைத் தெரியும்.

இவ்வாக்கியம் 'ராஜாவுக்கு இராமனைக் காட்டிலும் பணம் அதிகம் உள்ள ஒருவரைத் தெரியும்' என்றும் 'ராஜாவுக்கு இராமனைத் தெரிந்திருப்பதை விடப் பணக்காரனை அதிகம் தெரியும்' என்றும் பொருள் மயக்கம் காட்டும். இதுபோன்று பின்வரும் வாக்கியமும் அமைப்புப் பொருண்மை மயக்கம் காட்டும்.

இராமன் தனது தாயை நேசிப்பது போன்று கண்ணனும் நேசிக்கின்றான்

மேற்கண்ட வாக்கியம் 'இராமன் தனது தாயை நேசிக்கிறான். அது போல் கண்ணனும் தன் தாயை நேசிக்கின்றான்' என்றும் 'இராமன் தனது தாயை நேசிக்கும் அளவுக்கு கண்ணனும் இராமனது தாயை நேசிக்கின்றான்' என்றும் பொருண்மை மயக்கம் காட்டும். மேற்சொன்ன வாக்கியம் உண்மையில் பொருண்மை மயக்கம் காட்டுகின்றதா? என்ற கேள்வி எழலாம். புற வாக்கியத்தின் இரண்டாவது பகுதியில் செயப்படுபொருள் வெளிப்படையாக இல்லாததால் இவ்வாக்கியம் பொருண்மை மயக்கத்தைக் காட்டுகின்றது. எனவே பொருண்மை மயக்கம் என்பது தோற்றத்தைத் தருவது; அதை பொருண்மைசார் குறை வரையறை (underdetermination) என்று விளக்கலாம்.

பொருண்மை மயக்கம் என்பது அடிப்படையில் மொழி வெளிப்பாடுகளின் உடைமைப்பண்பு என்றாலும் மக்கள் எவ்வாறு மொழியைப் பயன்படுத்துகின்றனர் என்பதன் அடிப்படையிலும் பொருண்மை மயக்கம் காட்டுகின்றனர் என்றும் கூறப்படுகின்றது. அவர்களின் சொற்கள் பொருண்மை மயக்கம் இன்றி இருந்தாலும் இது நேரலாம்; அவர்கள் பயன்பத்துகின்ற சொற்கள் அவர்கள் விரும்பும் பொருளை உணர்த்தாதும் போகலாம். கண்டிப்புடன் கூறினால் பொருண்மை மயக்கம் என்பது பேசுபவரின் பொருண்மையை உட்படுத்தாமல் மொழிப் பொருண்மையை உட்படுத்துகின்றது. பொதுவாக ஒருவர் பொருண்மை மயக்கம் காட்டும் சொற்களையோ தொடர்களையோ வாக்கியங்களையோ பயன்படுத்தினால் அவர் உணர்வோடு (மனதார) தான் விருப்பாதாத பொருண்மைகளை வெளிப்படுத்துவதில்லை. இருப்பினும் ஒருவர் பொருண்மை மயக்கம் உள்ள சொற்களைக் கேட்கும் போது உடனடியாக அதைப் புரிந்துகொண்டு அதன் பொருந்தாத அர்த்தங்களை விட்டுவிடுவார். மக்கள் பொருண்மை மயக்கம் தரும் மொழியைப் பயன்படுத்தும் போது பொதுவாக அதன் பொருண்மை மயக்கம் உத்தேசிக்கப்படுவதில்லை. இருப்பினும் சில வேளைகளில் பொருண்மை மயக்கம் வேண்டுமென்றே செய்யப்படும்.

2.2.2.1 அமைப்புப் பொருண்மை மயக்கத்தின் வகைகள்

சொற்பொருண்மை மயக்கம் தனிச் சொற்களை ஆய்வது மற்றும் அவற்றின் பொருண்மைகளை மாற்றுவது ஆகிய சிக்கல்களை உள்ளடக்கும். அமைப்புப் பொருண்மை மயக்கம் தொடரியல் அமைப்புகள் மற்றும் வாக்கியங்களின் உருப்படுத்தங்கள் ஆகிய சிக்கல்களை உள்ளடக்கும். ஒரு ஒழுங்குமுறையில் பயன் படுத்தப்பட்டுள்ள இலக்கணத்தின் அடிப்படையில் ஒரு வாக்கியத்தின் அக அமைப்பை ஒன்றிற்கும் மேற்பட்ட வழிகளில் ஆய்வது பொருண்மை மயக்கத்திற்கு வழிவகுக்கும். 'ஒரு வாக்கியத்தின்' மற்றும் 'ஒழுங்குமுறையில் பயன்படுத்தப்பட்டுள்ள இலக்கணத்தின் அடிப்படையில்' என்ற சிறப்புப்பண்புகள் முக்கியமானவைகள் ஆகும். பொரும்பாலான மொழிபெயர்ப்பு ஒழுங்குமுறைகள் ஒரு வாக்கியத்தின் வரிசைமுறை ஆய்வுடன் எல்லைப் படுத்தப்பட்டுள்ளது; அவை பத்தி போன்ற மொழிபெயர்ப்பின் பெரிய அலகுகளை பொதுவாகக் கையாளுவதில்லை; இருப்பினும் வாக்கியங்களைத் தொடர்புபடுத்தும் படிக்கு பண்புக்கூறுகளைக் கையாளுவதற்குச் சில முயற்சிகள் எடுக்கப்பட்டுள்ளன. எடுத்துக்காட்டாக மாற்றுப் பெயர்கள், தீம்-ரீம் (theme-reem) அமைப்புகள் மற்றும் பிற. பெரும்பாலான இம்மாதிரியான ஒழுங்கமைப்புகள் யாவும் பரிசோதனை அளவிலேயே உள்ளன. எந்தப் பகுப்பானும் செயல்முறை படுத்தப்பட்டுள்ள இலக்கணத்தின் எல்லையைத் தாண்டி செல்ல இயலாது என்பது இரண்டாவது சிறப்புப்பண்பின் நினைவூட்டியாகும். ஒரு மனிதவினப் படிப்பவர் செய்ய இயலும் வேறுபாடுகளை ஒரு இலக்கணம் செய்ய இயலாவிடில் பகுப்பான் மாறுபட்ட பகுப்பாய்வுகளுக்கு இடையில் தீர்மானம் எடுக்க இயலாது. பின்வரும் பகுதியில் பல எடுத்துக்காட்டுகள் தரப்பட்டுள்ளன. இலக்கணம்தான் ஒரு குறிப்பிட்ட அமைப்புக்கு ஒன்றுக்கும் மேற்பட்ட 'சட்டபூர்வமான' பொருள்கோண்மைகள் இருக்கின்றனவா என்றும் அப்படி இருந்தால் அது பொருண்மை மயக்கம் உள்ளது என்றும் தீர்மானிக்கின்றது. பல பொருள்கோள்கள் ஒரு மனிதன் கண்டுபிடிக்கவியலும் 'உண்மையான' பொருண்மை மயக்கங்களுக்கும் மனிதப் படிப்பவர் தெரிந்துகொள்ளத் தேவையில்லாத 'ஒழுங்குமுறை' பொருண்மை மயக்கங்களுக்கும் வேறுபாடு காண்பது ஏற்கத்தக்கதாகும்.

2.2.2.1.1 உண்மையான அமைப்புப் பொருண்மை மயக்கம்

மொழியியலார்கள் பொதுவாக முறையான பகுப்பாய்வால் வெளிப்படுத்தப்படும் மாறுபட்ட தொடரியல் பொருள்கோள்களை வெளிச்சம் போட்டுக்காட்ட உண்மையான அமைப்புப் பொருண்மைகள் பற்றி எழுதுவதை விரும்புவார்கள். பின்வரும் எடுத்துக்காட்டுகளை எடுத்துக்கொள்வோம்.

39. பழைய மாணவர் விடுதியில் அவன் தங்குகின்றான்.

40. மீன் விழுங்கிக் குழந்தை இறந்தது.

41. சின்னசாமி காமராசர் பல்கலைக்கழகத்தில் படிக்கிறான்.

இவ்வாக்கியங்களைப் படிப்பவர் ஒவ்வொரு வாக்கியத்திற்கும் கீழே தரப்பட்டுள்ளது போல் ஒன்றிற்கும் மேற்பட்ட பொருள்களைப் பெற இயலும்.

39அ. மாணவர்களுக்கான பழைய விடுதியில் அவன் தங்குகின்றான்.

39ஆ. பழைய மாணவர்களுக்கான விடுதியியில் அவன் தங்குகின்றான்.

40அ. மீனை விழுங்கியதால் குழந்தை இறந்தது.

40ஆ. குழந்தையை மீன் விழுங்கியதால் குழந்தை இறந்தது.

41அ. 'காமராசர் பல்கலைக்கழகத்தில்' சின்னசாமி படிக்கிறான்.

41ஆ. 'சின்னசாமி காமராசர் பல்கலைக்கழத்தில்' அவன் படிக்கிறான்.

இந்த வாக்கியங்களில் பொருண்மை மயக்கங்களைப் புரிந்துகொள்வது எப்பொழுதும் எளிது அல்ல. ஒரு சூழலில், ஒரு நிலைமையில் அல்லது ஒரு குறிப்பிட்ட பனுவலில் வாக்கியங்கள் அதிக மயக்கம் காட்டலாம். எடுத்துக்காட்டாக (40)-ஆவது வாக்கியம் ஒரு கதையில் வந்தால் கதையிலிருந்து குழந்தை மீன் விழுங்கியதா? குழந்தையை மீன் விழுங்கியதா? என்று தெளிவாகத் தெரியும். இருப்பினும் இயந்திர மொழி பெயர்ப்பு ஒழுங்கமைப்புகள் இந்தச் சூழல் தடயங்களை மிகக் குறைந்த அளவிலன்றிப் பயன்படுத்த இயலாது. இந்த நேர்வுகளில் 'ஒழுங்குமுறை' பொருண்மை மயக்கம் 'உண்மை' பொருண்மை மயக்கத்துடன் பொருந்தும்.

2.2.2.1.2 யதேட்சை அமைப்புப் பொருண்மை மயக்கம்

இருப்பினும் பிற வினைகள், பெயர்கள், முன்னுருபுகள் இவற்றால் இடம்பெயர்ப்பதலால் இம்மாதிரியான எடுத்துக்காட்டுகளின் பொதுமையாக்கம் படிக்கும் ஒருவரால் பொருண்மை மயக்கம் உள்ளதாகக் கருதப்பட்டாத மேலும் கூடுதல் பொருண்மை மயக்கங்களுக்குக் கொண்டுசெல்லும். பொருண்மை மயக்கத்தை நீக்கப் போதுமான தகவலைக் கொண்டிராததால் ஒழுங்குமுறை உண்மையான பொருண்மை மயக்கத்தையும் ஒழுங்குமுறை பொருண்மை (system ambiguities) மயக்கத்தையும் ஒன்றுபோல் கையாளும். அம்மாதிரியன நேர்வுகள் யதேட்சை (அல்லது 'ஒழுங்குமுறை') அமைப்புப் பொருண்மை மயக்கம் (accidental structural ambiguities) என்று அழைக்கப்படும்; இவ்வழியில் தவறான பொருள்கோளை நிராகரிக்க எந்தத் தகவல் இலக்கணத்தில் சேர்க்கப்படவேண்டும் என்று மொழியிலார் முயற்சிக்கலாம் என்பது இவற்றை ஆய்வதற்கான பகுதி ஆர்வம் ஆகும்.

வகைப்பாட்டு மயக்கத்தைக் கொண்ட சொற்களின் ஒன்றிணைப்பின் காரணமாக, தொடரியல் உறுப்புகளின் மாற்றுப் பயன்பாடு காரணமாக, அல்லது தொடரியல் உறுப்புகளின்

சாத்தியமான வேறுபட்ட ஒன்றிணைபு காரணமாக யதேட்சையான அமைப்புப் பொருண்மை மயக்கம் ஏற்படும். அமைப்புப் பொருண்மை மயக்கத்தின் வகைகள் மொழிக்கு மொழி வேறுபடும், முக்கியமாக இலக்கணத்திற்கு இலக்கணம் வேறுபடும். இருப்பினும் ஓரளவுக்கு விரிந்த செயலெல்லை உள்ள இலக்கணம் உருவாக்கும் பொருண்மை மயக்கங்களின் வகைகளை எடுத்துக்காட்டுகளாக விவாதிப்பது பயனுள்ளதாய் அமையும்.

ஒரு சொல் ஒரே தொடரியல் சூழலில் வேறுபட்ட செயல்பாட்டில் வரும் என்ற உண்மையின் காரணமாகப் பல பொருண்மை மயக்கங்கள் ஏற்படும். இது முன்னர் விவாதிக்கப்பட்ட வகைப்பட்டு பொருண்மை மயக்கத்தின் சாத்தியமான பின்விளைவு ஆகும். வாக்கியம் 39-இல் *பழைய* என்பது மாணவர்களுக்கும் விடுதிக்கும் பெயரடையாக வர இயலுமாகையால் பொருண்மை மயக்கம் ஏற்பட்டது. வாக்கியம் 40-இல் *மீன்* என்பது *விழுங்கு* என்ற வினையின் எழுவாய்ப் பெயராகவும் செயப்படுபொருள் பெயராகவும் வர இயலுமாகையால் பொருண்மை மயக்கம் ஏற்பட்டது. இந்தப் பொருண்மை மயக்கங்கள் ஒரே வாக்கியங்களின் வேறுபட்ட அமைப்புப் பொருள்கோளில் பிரதிபலிக்கின்றது. சாம்ஸ்கி இலக்கண மாதிரி அடிப்படையில் ஒரே 'புற அமைப்புக்கு' (surface structure) வேறுபட்ட அக அமைப்புகள் (deep structures) இருப்பதன் காரணமாக அவை அக அமைப்புப் பொருண்மை மயக்கங்கள் எனப்படும்.

42. உண்ணும் போது நான் அவளைப் பார்த்தேன்.

மேற்கண்ட வாக்கியத்தில் *நான்* என்பதும் *அவள்* என்பதும் *உண்* என்பதன் எழுவாயாக வர இயலும் என்பதால் இரு பொருள்கோள்கள் சாத்தியமாகும்.

42அ. [நான் உண்ணும் போது] நான் அவளைப் பார்த்தேன்.

42ஆ. [அவள் உண்ணும் போது] நான் அவளைப் பார்த்தேன்.

பின்வரும் வாக்கியத்தைக் கருத்தில் கொள்ளவும்.

43. நான் பணத்துடன் அவளைப் பார்த்தேன்

மேற்கண்ட வாக்கியத்தில் *பணம்* என்பது *நான்* என்பதற்கும் *அவள்* என்பதற்கும் உடனிருப்புப் பொருளாக வருமாகையால் அது பின்வருமாறு பொருண்மை மயக்கம் காட்டும்.

43அ. நான் பணம் வைத்திருக்கிற அவளைப் பார்த்தேன்.

43ஆ. நான் பணம் வைத்துக்கொண்டு அவளைப் பார்த்தேன்.

பின்வரும் வாக்கியத்தில் வினைப்பெயர் (வினை+காலம்+அது) வினைச் செயலையும் அதன் செயப்படுபொருளையும் உணர்த்தி பொருண்மை மயக்கத்துடன் வரும். பின்வரும் வாக்கியத்தைக் கருத்தில் கொள்ளவும்.

44. அவள் பார்த்ததை நான் பார்த்தேன்.

இதைப் பின்வருமாறு இரு விதமாகப் பொருள்கோள் செய்யலாம்.

44அ. நான் அவள் பார்த்ததாகியச் செய்யலைப் பார்த்தேன்.

44ஆ. நான் அவள் எதைப் பார்த்தாளோ அதைப் பார்த்தேன்.

இது போலவே பின்வரும் வாக்கியங்களும் பொருண்மை மயக்கம் தருவனவாகும்.

45. அவர்கள் கேட்க முடியாததை ஆசிரியரிடம் கூறினர்.

இவ்வாக்கியம் பின்வருமாறு இருவிதமாகப் பொருள்படும்.

45அ. அவர்கள் ஆசிரியர் கூறுவதைக் கேட்க இயலவில்லை என்று கூறினர்.

45ஆ. அவர்கள் ஆசிரியரிடம் தாங்கள் எதைக் கேட்க இயலவில்லையோ அதைக் கூறினர்.

பின்வரும் வாக்கியத்தில் இரண்டாவது முதன்மை வினைமுற்றுத் தொடரில் வரும் செயல் எச்சத்தொடரிரில் வரும் செயலின் உடனிகழ்வுச் செயலாகவோ அல்லாதோ வரலாம்.

46. அவன் இருந்த இடத்தில் அவள் இருந்தாள்.

இதை இருவிதமாகப் பொருள் கொள்ளலாம்.

46அ. அவன் ஒரு இடத்தில் இருந்தான். அவன் எழுந்து போனபின் அதே இடத்தில் அவள் இருந்தாள்.

46ஆ. அவன் ஒரு இடத்தில் இருந்தான். அவன் கூடவே அவளும் இருந்தாள்.

பின்வரும் வாக்கியங்களில் 'விமான விபத்து' இருவிதமாகப் பொருள்கொள்ளப்படும்.

47. நீ விமான விபத்தைப் பற்றி செய்தித்தாளில் படித்தாயா?

48. நீ விமான விபத்தை நேரில் பார்த்தாயா.

முதல் வாக்கியத்தில் 'விமான விபத்து' செய்தியையும் இரண்டாவது வாக்கியத்தில் நிகழ்ச்சியையும் குறிக்கும்.

பின்வரும் வாக்கியத்தைக் கருத்தில் கொள்ளவும்.

49. ராஜா கண்ணனுக்குப் பணம் கொடுப்பதை நான் விரும்பவில்லை.

இவ்வாக்கியம் கவனக்குவிப்பு அடிப்படையில் பின்வருமாறு குறைந்தது மூன்றுவிதமாகப் பொருள்கோள் செய்யப்படும்.

49அ. ராஜா கண்ணனுக்குக் கொடுப்பதை நான் விருப்பவில்லை. (கண்ணனைத் தவிர வேறு யாருக்காவது கொடுப்பதை விருப்பியிருப்பேன்.)

49ஆ. ராஜா பணம் கொடுப்பதை நான் விரும்பவில்லை. (பணம் தவிர வேறு எதையாவது கொடுப்பதை விருப்பியிருப்பேன்.)

49இ. ராஜா கண்ணனுக்குக் கொடுப்பதையும் பணம் கொடுப்பதையும் நான் விரும்பவில்லை

விருப்புவதாகிய செயல் *கண்ணன்* என்பதைக் கவனக்குவிப்பு செய்தோ *பணம்* என்பதைக் கவனக்குவிப்பு செய்தோ அல்லது இரண்டையும் கவனக்குவிப்பு செய்தோ அமையலாம்.

2.2.2.2 அமைப்புப் பொருண்மை மயக்கத்தை நீக்குதல்

தொடரியல் ஆய்வு ஒன்றுக்கும் மேற்பட்ட சாத்தியமான பொருள்கோளைத் தந்தால் சரியான ஒன்றைத் தேர்ந்தெடுப்பது அவசியமாகும். பெரும்பாலும் பொருள்கோளின் தேர்வு அடிப்படையில் இலக்கு மொழியின் மொழிபெயர்ப்பு அமையும் என்பது இதன் காரணமாகும். எடுத்துக்காட்டாக

50. நான் அவள் பர்த்ததைப் பார்த்தேன்.

என்ற வாக்கியம் இருவித பொருள்கோள் அடிப்படையில் இருவிதமாக மொழிபெயர்ப்பு செய்யப்படும்.

50அ. I saw her seeing me.

50ஆ. I saw what she saw.

அமைப்புப் பொருண்மை மயக்கத்தை நீக்குவதற்குப் பல வழிகள் உள்ளன: பொருண்மையியல் அல்லது பிற மொழியியல் தகவல்களைப் பயன்படுத்தல், சூழல் தடயங்களைப் பயன்படுத்தல், 'உண்மை உலக அறிவைப்' பயன்படுத்தல், ஊடாடிக் கருத்தறிதல். 'சரியான ஊகம்' என்ற வழுநிலை உத்தி அல்லது 'சுதந்திர அனுமானத்தை' எதிர்பார்ப்பது இவற்றைப் பயன்படுத்திப் பொருண்மை மயக்கத்தைத் தவிர்ப்பது என்பன பிற விருப்பத்தேர்வுகளாகும்.

2.2.2.2.1 மொழியியல் அறிவைப் பயன்படுத்தல்

பெரும்பாலும் சாத்தியமான பொருண்மை மயக்க வாக்கியங்கள் மொழியியல் அறிவு என்று அழைக்கப்படுகின்றதைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்யப்படலாம். பல வகையான மொழியியல் அறிவுகள் உள்ளன; ஆனால் அவை எல்லாம் வாக்கியங்கள் விளக்கும் உண்மை வாழ்க்கை நிகழ்வுகளுக்குப் பதிலாகச் சொற்களையும் அவை ஒன்று சேரும் வழி பற்றிய தகவல்களையும் பயன்படுத்துகின்றன என்பது அவைகளுக்கிடையே உள்ள பொதுவான தன்மையாகும்.

ஒரு வழிமுறை சேர்ந்துவருகை கட்டுப்பாடுகள் பற்றிய தகவலை அதாவது ஒரு அமைப்பில் சில குறிப்பிட்ட உறுப்புகளின் இருப்பு பிற உறுப்புகளின் இருப்பைப் பாதிக்கின்றன என்ற அறிகுறிகளைப் பகுப்பானுக்குத் தருவதாகும். தெளிவான எடுத்துக்காட்டு வினைகளுக்கு துணைவகைப்பாட்டுச் சட்டங்களைப் (subcategorization) பயன்படுத்துவதாகும். இவை ஒரு குறிப்பிட்ட வினை எவ்வகையிலான பங்கெடுப்பாளர்களை ஏற்கும் என்பதைக் காட்டும். எடுத்துக்காட்டாக, *கொடு* என்ற வினை கொடுப்பவரைக் குறிக்கும் ஒரு பெயரை எழுவாயாகவும் (subject) கொடுக்கப்படும் பொருளைக் குறிக்கும் ஒரு பெயரை நேரடி செயப்படு பொருளாகவும் (direct object) பெறுபவரைக் குறிக்கும் பெயரை மறைமுகச் செயப்படு பொருளாகவும் (indirect object) ஏற்கும். மேலும் பெயர்களுக்கு பொருண்மைப் பண்புக்கூறுகளைத் தந்து இந்தத் தொடரியல் பங்களிப்புகளை எவ்வகையிலான பெயர்கள் நிரப்பும் என்பதை ஓரளவுக்குக் குறிப்பிட இயலும். எடுத்துக்காட்டாக, கொடுப்பவர் விலங்கினத்தைச் (மனிதவினம் உள்ளடக்கிய) சார்ந்ததாக இருக்கவேண்டும் போன்றவை. இவ்வாறு சாத்தியமான பொருண்மை மயக்கம் உள்ள இரு வாக்கிய இணைகள் தரப்படுகையில் பகுப்பான் சரியான பொருள்கோளைச் செய்யும்.

51. நான் புத்தகத்தைப் படித்தேன்.

52. நான் புத்தகத்தைக் கிழித்தேன்.

முதல் வாக்கியத்தில் *படி* என்ற வினை படிக்கக்கூடிய பொருளைச் செயப்படுபொருளாக ஏற்கும் ஆகையால் *புத்தகம்* என்பது அதில் உள்ள செய்தியைக் குறித்து நிற்கும். இரண்டாவது வாக்கியத்தில் கிழி என்ற வினை கிழிக்கக்கூடிய பொருளைச் செயப்படுபொருளாக ஏற்கும் ஆகையால் *புத்தகம்* என்பது அதன் பாகமாகிய தாள்களைக் குறித்து நிற்கும்.

இம்மாதிரியான தகவலைக் கூடுதல் பொதுவான நிலையில் கையாள இயலும்; அதாவது இணைதிறன் (valency) மற்றும் வேற்றுமை இலக்கணம் (case grammar) அடிப்படையில். இணைதிறனில் வினைகள் அவர்கள் ஏற்கும் பங்கெடுப்பாளர்களின் எண்ணிக்கை மற்றும் வகை அடிப்படையில் பண்பாக்கம் செய்யப்படும். வேற்றுமை இலக்கணத்தில் சார்ந்திருக்கும் பங்கெடுப்பாளர்களின் பங்களிப்பு அடையாளம் காணப்படும்; எடுத்துக்காட்டாக செயலி (agent), தாங்கி (patient)/ செயப்படுபொருள் (object), கருவி (instrument), முறை (manner), உடன்வருபவர்/கூட்டாளி (accompanier), இருப்பிடம் (location) என்பன சில பங்களிப்புகளாகும். ஒரு எடுத்துக்காட்டான பொதுமையாக்கம் செயப்படுபொருள் குன்றா வினைகளின் எழுவாய்கள் செயலிகளாகும்; அவைகள் எடுத்துக்காட்டாக 'விலங்கினத்தன்மை' (animate), 'திறன்' (potent) போன்ற பண்புகளைக் கொண்டிருக்கும். எடுத்துக்காட்டாகக் *கொடு* என்ற வினை செயலி,

பெறுபவர் (receipient), தாங்கி (patient) என்ற வேற்றுமைப் பங்களிப்புகளை எதிர்பார்க்கக்கூடும். பின்வரும் வாக்கியங்களைக் கருத்தில் கொள்ளவும்.

53. அவன் அவளுடன் போரிட்டான்.

54.அவன் அவளுடன் வந்தான்.

முதல் வாக்கியத்தில் *அவள்* தாக்கி பங்களிப்பையும் இரண்டாவது வாக்கியத்தில் *அவள்* உடன்வருபவர் பங்களிப்பையும் செய்கின்றது.

2.2.2.2.2 சூழல் அறிவு

நடைமுறையில் ஒரு சில வாக்கியங்கள் தாம் பொருண்மை மயக்கம் தருவன; பொருண்மை மயக்கநீக்கத்திற்கு எதுவும் உதவவில்லை என்றாலும் அவ்வாக்கியம் நிகழும் சூழல் எந்த பொருள்கோள் சரியானது என்ற அனுமானத்தைத் தரும். 40-ஆவது வாக்கியதைப் பற்றி விவாதிக்கையில் கதையின் போக்கு எது எதை விழுங்கியது என்பதை அனுமானிக்கும் என்று பார்த்தோம். அது முந்தைய வாக்கியத்திலோ பத்தியிலோ இயல்களிலோ கூறப்பட்டிருக்கலாம். இருப்பினும் ஒரு குறிப்பிட்ட நேர்வின் பொருண்மை மயக்கத்தை நீக்க உதவும் அறிவைப் பெற எங்கு பார்க்கவேண்டும் என்ற திட்டவட்டமான விதிகள் இல்லை என்பதால் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் சூழல் அறிவைப் பயன்படுத்த இயலாது. 'மீன் குழந்தையை விழுங்கியது' போன்ற பனுவலில் இருந்து பெறப்பட்ட அறிவைச் சேகரித்துவைக்கும் திறமையான வழிகள் இருக்கிறது என்று வைத்துக்கொண்டாலும் எந்த அறிவின் பாகம் பின்னர் பயனுள்ளதாக இருக்கும் அல்லது எவ்வளவு காலம் அவை சேகரித்துவைக்கப்படவேண்டும் என்பதை அறிவது கடினம். ஏனென்றால் தரப்பட்ட ஒரு பனுவலின் ஒவ்வொரு வாக்கியத்திலிருந்தும் அனுமானிக்கப்படும் ஒவ்வொரு உண்மைகளையும் பிரித்தெடுப்பதும் சேகரித்துவைப்பதும் நடைமுறைக்கு ஒவ்வாததாகும்.

2.2.2.2.3 உண்மையான உலக அறிவு

தொடரியல் ஆய்வு போதுமானதாக இல்லாதிருக்கும் போது பொருண்மை மயக்க நீக்கத்திற்கு மூன்றாவது அணுகுமுறை உண்மை உலக அறிவு என்று பொதுவாக அழைக்கப்படும் மூலவளத்தைக் கொண்டிருப்பதாகும். உண்மை உலக அறிவைப் பயன்படுத்தி அமைப்புப் பொருண்மை மயக்கத்தை நீக்கும் எடுத்துக்காட்டு கீழே தரப்பட்டுள்ளது.

55. நான் பணத்துடன் புலியைப் பார்த்தேன்.

உண்மை உலக அறிவுக்கும் மொழியியல் அறிவுக்கும் இடையில் உள்ள வேறுபடுத்தும் கோடு தெளிவானதல்ல. எடுத்துக்காட்டாக பின்வரும் வாக்கியங்களைக் கருத்தில் கொள்ளவும்.

55. நான் நேற்று நீ சொன்ன நபரை பார்ப்பேன்.

56. நான் நாளை நீ சொன்ன நபரைப் பார்ப்பேன்.

முதல் வாக்கியத்தில் *நேற்று* என்பது *சொல்* என்ற வினையின் வினையடையாக வருகின்றது; இரண்டாவது வாக்கியத்தில் *நாளை* என்பது *பார்* என்ற வினையின் வினையடையாக உள்ளது. *நேற்று* என்பது இறந்தகாலத்துடனும் *நாளை* என்பது எதிர்காலத்துடனும் உடன்பாடு காட்டும் என்பது மொழியியல் அனுமானமாகும். இருப்பினும் ஒருவர் தற்போது ஒன்றைச் செய்யவிருக்கும் திட்டத்தை இறந்தகாலத்தில் கூற இயலாது மற்றும் ஒருவர் ஒன்றைச் செய்துவிட்டதை எதிர்காலத்தில் கூறவியலாது என்பன உண்மை உலக அறிவாகும். வேறுவிதத்தில் கூறினால் மொழியியல் அறிவு பெரும்பாலும் உண்மை உலக அறிவுடன் ஒத்திருக்கும்.

தற்போது ஒரு குறிப்பிட்ட ஒழுங்குமுறையில் எல்லா பொருண்மை மயக்கங்களையும் குறியனாக்கம் செய்யவும் உட்படுத்தம் செய்யவும் நடைமுறையில் இயலாததாகும் என்பது இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளின் சிக்கல் ஆகும்; ஒப்பீட்டு அடிப்படையில் சூழல்கள் மற்றும் பயன்பாடுகளின் குறுகிய செயல்பரப்புக்கு எல்லைப்படுத்திக்கொண்ட ஒழுங்குமுறைகளுகும் இது இயலாததாகும். செயற்கை அறிவு மற்றும் கணிக்கும் தொழில் நுட்பம் இவற்றின் முன்னேற்றத்திற்குப் பின்னரும் இந்த நிலை எதிர்காலத்தில் மேம்படுவது இயலாததாகும். உலக அறிவின் கலவைத்தன்மையும் பெற இயலாமையும் விரைவான தீர்வுகளுக்குத் தடையாக உள்ளன.

2.2.2.2.4 பிற உபாயங்கள்

தீர்க்கப்படாத பொருண்மை மயக்கங்களின் எச்சம் (உண்மை பொருண்மை மயக்கமோ 'ஒழுங்குமுறை' பொருண்மை மயக்கமோ) எப்போதும் இருந்துகொண்டே இருக்கும். இச்சூழலில் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை மனித மொழிபெயர்ப்பாளர்கள் பயன்படுத்தும் உபாயங்களைக் கையாளக்கூடும். போட்டியிடுகிற அமைப்புகளில் எவை அதிகமாக எளிதில் புரிந்துகொள்ளப்படும் என்ற கேள்விகளுக்குப் போதுமான அளவு உளமொழியியல் ஆய்வுகள் உள்ளன.

சாத்தியம் என்றால் படைப்பாளியைக் கேட்பது இரண்டாவது உபாயம். சில ஊடாட்ட இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் (Interactive Machine Translation Systems) இந்த அணுகுமுறையை எடுத்துக்கொள்கின்றன: அவை எடுத்துக்கொண்ட பொருளின் அவர்களது அறிவுடன் பொருந்தும் படைப்பாளியின் விருப்பம் பற்றி தங்கள் புரிந்துகொள்ளலின் ஆய்வைத் தெரிந்தெடுக்க மனித இயக்கிகளிடம் (human operators) கேட்கின்றனர்.

57. வெப்பம் சார்ந்த உமிழ்வால் அழுத்த ஆய்வின் நுட்பங்களின் புலனாய்வு

இது *'வெப்பம் சார் உமிழ்வால் ஆய்வு'* அல்லது *'வெப்பம் சார் உமிழ்வால் புலனாய்வு'* மற்றும் *'அழுத்தத்தின் ஆய்வுக்கு நுட்பம்'* அல்லது *'அழுத்தத்தின் நுட்பத்தின் ஆய்வு'* எனத் தீர்மானிக்க இந்த எடுத்துக்காட்டின் பொருள்கோள் சிறப்பான அறிவியல் அறிவை வேண்டும்.

மூன்றாவது உபாயம் ஒப்பீட்டு அடிப்படையில் ஒன்றுக்குப் பதிலாக ஒன்றின் சாத்தியத்தியம், உட்படுத்தப்படும் சொற்களைக் கருத்தில் கொள்ளாமல் எந்த அமைப்பு ஏறக்குறைய சரியானது என்பதன் அடிப்படையில் நல் ஊகத்தைச் செய்வது. எடுத்துக்காட்டாக மேற்சொன்னது போன்ற கலவைத்தன்மையான பெயர்த் தொடர்களில் 'நல் ஊகம்' ஒவ்வொரு வேற்றுமைத் தொடரும் அதன் பின்னால் வரும் பெயரை அடை செய்யும்; அல்லது வினையடைத் தொடர் எப்போதும் மிக சமீபத்திய வினையை அடைசெய்யும். வெளிப்படையாக இந்த குருட்டுத்தனமான அணுகுமுறை சில வேளைகளில் தவறான ஆய்வுக்கு வழிவகுக்கும்; ஆனால் இந்த ஊகங்கள் சரியான நோக்கம் உடையதாய் இருந்தால் சரியான ஆய்வை அதிக நஷ்டமில்லாமல் பெரும்பாலும் செய்ய இயலும்.

இறுதியாக பொருண்மை மயக்கம் நீக்கப்படவேண்டியதில்லை ஏனென்றால் அதை இலக்கு மொழியில் தக்கவைக்க இயலும். இந்த விருப்பத்தேர்வு ஒற்றுமையுள்ள அமைப்பு மற்றும் சொற்றொகை உள்ள மொழிகளுக்கு இருக்கும். எடுத்துக்காட்டாக பின்வரும் வகையிலான பொருண்மை மயக்கம் தமிழில் இல்லை ஆனால் ஜெர்மன் மொழியில் உள்ளது.

58.The man saw the girl with the telescope.

ஒரு முக்கியமான செய்தி அழுத்தமாகக் கூறப்படவேண்டும். வேறுபட்ட பொருண்மை மயக்கங்களின் வேறுபாடுகளைச் செய்வது ஒழுங்குமுறையின் பார்வை அடிப்படையில் தேவையற்றதாகும். பொருண்மை மயக்கம் மொழியியல் அறிவை, சூழல் அறிவை அல்லது உண்மை அறிவை வேண்டுமா என்பது முக்கியமல்ல; மாறாகப் பொருண்மை மயக்கம் நீக்குவதை அனுமதிக்கப் பொருத்தமான தரவுகள் கிடைக்கிறதா என்பது முக்கியமாகும். ஒழுங்குமுறை ஒரு பொருண்மை மயக்கம் இருக்கின்றது என்று அறிந்துகொண்டால் அதை நீக்குவதற்கு அதற்கு வழிவகைகள் இருக்கும்; ஒழுங்குமுறை அதை நீக்குவதைத் தொடரும்.

2.2.3 குறிப்புப்பொருள் பொருண்மை மயக்கம் (referential ambiguity)

பேசுபவர் ஒரு குறிப்பிடு பொருளைக் குறிப்பிட ஒரு குறிப்பிடும் வெளிப்பாட்டைப் பயன்படுத்துகையில் கேட்பவர் மற்றொரு குறிப்பிடு பொருளை எண்ணும் போது இவ்வகையிலான பொருண்மை மயக்கம் ஏற்படும். பின்வரும் எடுத்துக்காட்டைக் கருத்தில் கொள்ளவும்.

59. ராஜா எனது நண்பன்.

கேட்பவர் ராஜா என்ற நபரை வேறு ஒரு நபராகப் புரிந்துகொள்ளக்கூடும். முற்சுட்டை (anophora) ஒன்றிற்கும் மேற்பட்ட சுதந்திர வடிவங்களுடன் தொடர்புபடுத்த இயலும் போது குறிப்பிடுபொருள் மயக்கம் (referential ambiguity) ஏற்படும். எடுத்துக்காட்டாக ஒரு மாற்றுப் பெயர் அதன் முன்வருகிளவியால் முன்தொடரப்பட்டிருக்கும்; அதிலிருந்து மாற்றுப்பெயர் அதன் பொருள்கோளைப் பெறும். இருப்பினும் சில அமைப்புகளில் நாம் ஒன்றுக்கும் மேற்பட்ட சுதந்திர வடிவங்கள் ஒரே மாற்றுப்பெயருக்கு முன்வருகிளவியாகச் செயல்படுவதைக் காணலாம். பின்வரும் வாக்கியத்தைக் கருத்தில் கொள்ளவும்.

60. ராஜா கண்ணனிடம் ராதை அவனுக்கு வேண்டிக் காத்திருப்பதாகக் கூறினான்.

மேற்சொன்ன வாக்கியத்தில் *அவன்* என்பது ராஜாவையோ கண்ணனையோ குறிப்பிட இயலும். இதை முற்சுட்டு மயக்கம் (anaphoric ambiguity) என்று குறிப்பிடலாம்.

2.2.4 நோக்கப் பொருண்மை மயக்கம் (Scope ambiguity)

ஒரு வாக்கியத்தில் ஒரு உறுப்பின் பொருண்மையை அதன் அமைப்புச் சூழலில் பிற உறுப்புகள் தீர்மானித்தால் அந்த உறுப்பு அதன் குறிப்பைத் தீர்மானிக்கும் உறுப்புகளின் நோக்கத்தில் இருக்கின்றது எனலாம். இவ்வாறு உறுப்புகள் அவை வரும் அமைப்புச் சூழல் அடிப்படையில் அவற்றின் பொருள்கோளில் வேறுபடக்கூடும். பின்வரும் எடுத்துக்காட்டில் மாற்றுப் பெயர் *அவள்* என்பதன் பொருள்கோளைக் கருத்தில் கொள்ளவும்.

61. ஒவ்வொரு இளவரசியும் அவள் ஒரு இளவரசனை கண்டுபிடிப்பாள் என்பதை அறிவார்கள்.

அவள் என்ற மாற்றுப் பெயர் ஒரு குறிப்பீட்டுப் பயன்பாட்டைக் (referential use) கொண்டிருக்கக்கூடும். அதாவது முன்னரே கூறப்பட்ட ஒருவரை அல்லது சுட்டிக்காட்டப்பட்ட ஒருவரை. அது ஒரு கட்டுண்ட மாற்றுப் பெயராகப் பயன்படுத்தப்படக்கூடும்; இந்நேர்வில் மாற்றுப் பெயர் *ஒவ்வொரு இளவரசிகள்* குறிப்பிடும் வெளிப்பாட்டால் கட்டுண்டுள்ளது; அந்த வெளிப்பாடு குறிப்பிடும் ஒவ்வொரு நபரையும் குறிப்பிடப் பயன்படுத்தப்படும். நோக்கப் பொருண்மை மயக்கத்தின் பிற நேர்வுகள் பின்வரும் வாக்கியங்களால் எடுத்துக்காட்டப்பட்டுள்ளது.

62. அவன் ஒரு கேள்விக்கு விடையளிக்கவில்லை

63. இரண்டு மாணவர்கள் ஒவ்வொரு ஆசிரியரிடமும் பேசினர்.

முன்வரும் எடுத்துக்காட்டுகள் ஒத்தறி/ஒப்பீட்டு நோக்கம் (relative scope) என்று அழைக்கப்படும்.

அளவைகளான *சில, எல்லா, ஒருவரும்* என்பன அவை காட்டும் நோக்கம் அல்லது பரப்பெல்லை அடிப்படையில் பொருண்மை மயக்கத்தை வெளிப்படுத்தும். எடுத்துக்காட்டாகப் பின்வரும் வாக்கியங்களைக் கருத்தில் கொள்ளவும்.

64. எல்லாப் பெண்களும் அந்தத் திரைப்படத்தை விரும்பவில்லை

என்ற வாக்கியம் 'எல்லாப் பெண்களும் அந்தத் திரைப்படத்தை விரும்பவில்லை; சிலர் விரும்புகிறார்கள்' என்றும் 'ஒரு பெண் கூட அந்தத் திரைப்படத்தை விரும்பவில்லை' என்றும் பொருள்கொள்ளப்படும்.

2.2.5 பொருண்மை மயக்கம் வேறுபடுத்தல்

உங்கள் சொற்கள் குறிப்பிடுவது 'நீங்கள் என்ன பொருளை உணர்த்துகிறீர்கள் என்பதைப் பொறுத்தது' என்பது வெற்றுரையாகும். ஒருவர் என்ன கூறுகின்றார் என்பதால் அவர் வேறுபட்ட விஷயங்களைப் பொருளுணர்த்த இயலும்; இது இதைச் செய்ய இயலும் வேறுபட்ட வழிகளைக் கூறவில்லை. பொருண்மை மயக்கம் என்பது ஒரு வழி; ஆனால் பிற பல உள்ளன: ஒப்புருமொழியம் (homonymy), தெளிவின்மை (vagueness), ஒப்பீடுத்தன்மை (relativity), சுட்டுமை (indexicality), உருவகத்தன்மை (nonliterality), மறைமுகத்தன்மை (indirection). இந்த எல்லா நிகழ்வுகளும் மொழிப் பொருண்மையின் பன்மடங்குத் தன்மையிலிருந்து வேறுபட்டவைகளை எடுத்துக்காட்டும்.

ஒரு வெளிப்பாடு வெளிப்புறமான நேர்வுகளை அனுமதித்தால் அது தெளிவற்றது எனப்படும். *'மொட்டையானது', 'கனமானது', 'பழையது'* என்பன தெளிவற்றவைகளுக்கு எடுத்துக்காட்டுகளாகும். அவைகள் அளவுகோலின் தெளிவற்ற இடங்களைக் குறிப்பிட பயன்படுத்தப்படுகின்றன என்பது அவற்றின் தெளிவின்மையை உணர்த்தும். *'புத்திசாலி', 'முட்டாள்', 'நேர்மை'* என்பன தெளிவற்றன ஆகும்; எனென்றால் பல காரணிகள் அடிப்படையில் அவற்றின் நேர்வுகள் தீர்மானிக்கப்படுகின்றன.

ஒப்பீடுத்தன்மை *'கனமான', 'வயதான'* என்ற சொற்களால் எடுத்துக்காட்டப்படும். கனமான மக்கள் கனமில்லாத யானைகளைவிட எடைகுறைவாக இருப்பர்; வயதான பூனை சில இளைஞர்களைவிட இளமையாக இருக்கக்கூடும். *ராஜா முடித்துவிட்டான், கண்ணன் தாமதிக்கிறான்* என்பன போன்ற வாக்கியங்கள் வேறுபட்ட ஒப்பீட்டுத் தன்மையைக் காட்டுகின்றன. *'முடி'* மற்றும் *'தாமதி'* என்பன பொருண்மை மயக்கம் காட்டுகின்றன என்று இது

காட்டவில்லை. அம்மாதிரியான வாக்கியங்கள் பொருண்மை அடிப்படையில் குறை வரையறை உள்ளவை.

'நீ', 'இங்கே', 'நாளை' என்பவைகளுக்கு அறுதியான பொருண்மை இருக்கின்றது; ஆனால் வேறுபடும் குறிப்பீட்டுப் பொருள்கள் இருக்கின்றன. எடுத்துக்காட்டாக *நாளை* என்பதன் பொருண்மை மாறுவதில்லை ஆனால் அது குறிப்பிடும் நாள் மாறும் (சுட்டுக்கள், சுட்டுமைகள்).

உருவகத்தன்மை (nonliterality), மறைமுகத்தன்மை (indirection), வெளிப்படையின்மை (inexpliciness) என்பன பேசுபவர் பொருள்படுவது அவரது சொற்கள் என்ன பொருள் தருகின்றன என்பதால் உறுதி செய்யப்படுவதில்லை (பேச்சுச் செயல்பாடுகள்). *'நீங்கள் அவனுக்குப் படி அளக்கின்றீர்கள்'*, 'நீங்கள் அவனுக்குக் காது *கொடுங்கள்'*, *'அவன் உங்களுக்குப் பூ சுற்றுகின்றான்'*, *'அந்த சினிமாவில் ஒன்றும் இல்லை'*, *'அவன் கூறுவதில் அர்த்தம் இல்லை'* என்பன மொழிப் பொருண்மை மயக்க நேர்வுகள் அல்ல; ஆனால் பேசுபவர்கள் பொருண்மை மயக்கம் காட்டுகின்றனர் என்று அடிக்கடி கூறப்படுவதால் இவை குழப்பதிற்கு உள்ளாக்கும்.

2.2.6 தத்துவ சம்பந்தம்

தத்துவ வேறுபாடுகள் கவனிக்கப்படாத பொருண்மை மயக்கம் காரணமாக மறைக்கப்படக்கூடும். எடுத்துக்காட்டாக ஒருவகையான பொருண்மை மயக்கம் இருக்கிறது; இது 'செயல்/பொருள்' (act/objet) அல்லது 'செயற்பாங்கு/விளைபொருள் (process/result) மயக்கம் என்று விளக்கப்படும். எடுத்துக்காட்டாக *'கட்டு'*, *'கடி'* என்பன இம்மயக்கத்தைக் காட்டும். மொழியின் தத்துவம் மற்றும் உள்ளம் இவற்றின் குழப்பம் 'அனுமானம்', 'கூற்று', 'எண்ணம்' இவற்றிற்கு இடையே உள்ள மயக்கத்தைப் புறக்கணிக்கும். மற்றொரு பொதுவான தத்துவப் பொருள் மயக்கம் வகை/நேர்வு (type/token) வேறுபாடு. அன்றாடச் சொற்களான *'விலங்கு'*, *'நூல்'*, *'ஊர்தி'* வகைகளுக்கும் நேர்வுகளுக்கும் பயன்படுத்தப்படுகின்றன. இது *'வாக்கியம்'*, *'சொல்'*, *'எழுத்து'* என்ற மொழியியல் சொற்களுக்கும் உண்மையாகும்; அதுபோல *'கருத்துரு'*, *'நிகழ்வு'*, *'உளநிலை'* என்ற தத்துவ இயல் சொற்களுக்கும் உண்மையாகும்.

2.3 முடிவுரை

இவ்வியலில் பொண்மை மயக்கங்கள் வேறுபட்ட வகைகளாகப் பிரிக்கப்பட்டு விளக்கப்பட்டுள்ளன. சொற்பொருண்மை மயக்கம் சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம், ஒப்புருமொழியம்சார் பொருண்மை மயக்கம், பல்பொருள் ஒருமொழியம்சார் பொருண்மை மயக்கம், மாற்றப் பொருண்மை மயக்கம் என வகைப்படுத்தப்பட்டு விளக்கப்பட்டுள்ளன.

அமைப்புப் பொருண்மை மயக்கம் உண்மையான அமைப்புப் பொருண்மை மயக்கம், யதேட்சையான அமைப்புப் பொருண்மை மயக்கம் என வகைப்படுத்தப்பட்டு விளக்கப்பட்டுள்ளன. அமைப்புப் பொருண்மை மயக்கத்தை நீக்குவதற்கான வழிமுறைகள் ஆயப்பட்டுள்ளன. பிற பொருண்மை மயக்கங்களான குறிப்பொருள் பொருண்மை மயக்கம், நோக்கப் பொருண்மை மயக்கம் போன்ற பிற பொருண்மை மயக்கங்களும் விளக்கப்பட்டுள்ளன.

இயல் 3

இயந்திர மொழி பெயர்ப்பும் சொற் பொருண்மை மயக்க நீக்கமும்

3.0 முன்னுரை

இவ்வியல் நான்சி இட் மற்றும் ஜீன் வெரோனிஸ் (Nancy Ide and Jean Veronis, 1998) என்போர் எழுதிய Words Sense Disambuation: The state of Art என்ற கட்டுரையைத் தழுவி அமைகின்றது. 50-களில் மொழியின் கணினி செயல்பாட்டின் ஆரம்ப நாட்களிலிருந்தே தானியக்கமாகச் சொற்பொருண்மை மயக்க நீக்கம் ஆர்வமுள்ளதாகவும் கருதத்தக்கதாகவும் இருந்தது. பொருண்மை மயக்க நீக்கம் ஒரு இடைப்பட்ட செயல்பாடு; அது அதன் ஒரு முடிவல்ல; இயற்கைமொழி ஆய்வுச் செயல்பாடுகளை அடைவதற்கு ஒரு மட்டத்திலோ பிற மட்டத்திலோ இது தேவையாகும். செய்தியைப் புரிந்துகொள்ளுதல், மனித-இயந்திரக் கருத்துப் பரிமாற்றம் போன்ற மொழிப் புரிதல் பயன்பாடுகளுக்கு இது வெளிப்படையான முக்கியத்துவம் வாய்ந்ததாகும். மொழி புரிந்துகொள்ளுதலை நோக்கமாகக் கொண்டிராத பயன்படுகளுக்கும் சில நேர்வுகளில் இது தேவை ஆகும்.

- இயந்திர மொழிபெயர்ப்பு (machine translation): பொருண்மை மயக்க நீக்கம் மயக்கமுள்ள சொற்களின் சூழல் அடிப்படையிலான அர்த்தங்களை அறிந்து கொள்வதற்கு உதவும்.
- தகவல் மீட்பு மற்றும் ஹைபர் டெக்ஸ்ட் நேவிகேசன் (information retrieval hyper text navigation): முக்கியச் சொற்களுக்கு வேண்டி தேடுதல் நிகழ்த்தப்பட்டால் சொல்லோ அல்லது சொற்களோ பொருத்தமில்லாதா ஆவணங்களில் வருவதை நீக்குவது விரும்பத்தக்கதாகும்; எடுத்துகாட்டாக, நீதி தொடர்பான குறிப்புகளுக்குத் தேடல் நிகழ்த்தப்பட்டால், *நீதிமன்றம்* (court) என்ற சொல்லை *சட்டம்* (law) என்பதுடன் தொடர்புப்படுத்துவதற்குப் பதிலாக அரசவையுடன் தொடர்புபடுத்தும் ஆவணங்களைக் களைய வேண்டும்.
- பொருளடக்கம் மற்றும் மையக் கருத்து ஆய்வு (content and thematic analysis): பொருளடக்கம் மற்றும் மையக் கருத்து ஆய்வுக்குப் பொதுவான அணுகுமுறை சொற்களின் அதாவது தரப்பட்ட கருத்துரு, கருத்து, மையக்கருத்து இவற்றைக் குறிப்பிடும் சொற்களின் முன் வரையறுக்கப்பட்ட வகைப்பாடுகளின் விநியோகத்தை ஒரு பனுவலைக் கடந்து ஆய்வதாகும். ஒரு சொல்லின் சரியான அர்த்தத்தில் மட்டும் அச்சொல்லின்

நேர்வுகளை உட்படுத்தவேண்டி இம்மாதிரியான ஆய்வுகளில் பொருண்மை மயக்க நீக்கத்தின் தேவை முன்னரே அறியப்பட்டுள்ளது.

- இலக்கண ஆய்வு (grammatical analysis): சொல் வகைப்பாடு அடையாளப்படுத்துவதற்குப் பொருண்மை மயக்க நீக்கம் பயனுள்ளதாக அமையும். முன்னுருபு இணைப்பு (prepositional attachment) போன்ற குறிப்பிட்ட தொடரியல் ஆய்வுகளுக்கும் சொற் பொருண்மை மயக்க நீக்கம் தேவையானதாகும் மற்றும் பொதுவாகப் போட்டியிடும் பகுப்புகளின் இடத்தைக் கட்டுப்படுத்தும்.
- பேச்சு ஆய்வு (speech processing): பேச்சு உருவாக்கத்தில் சொற்களின் சரியான ஒலியாக்கத்திற்குப் (phonetization) பொருண்மை மயக்க நீக்கம் தேவையாகும். எடுத்துக்காட்டாக *He conjured up an image* அல்லது *I conjure you to help me* என்பதிலுள்ள *conjure* என்ற சொல்லின் ஒலியாக்கத்திற்கு; மற்றும் பேச்சுப் புரிந்துகொள்கையில் சொல் பிரித்தலுக்கும் (word segmentation) ஓரோசைச் சொற்களை (homophone) வேறுபடுத்துவதற்கும் பொருண்மை மயக்க நீக்கம் தேவையாகும்.
- பனுவல் பகுப்பாய்வு (text processing): எழுத்துப்பிழை நீக்கத்திற்குச் சொற் பொருண்மை மயக்க நீக்கம் தேவையாகும். எடுத்துக்காட்டாக, எப்போது குறியீடுகள் இடவேண்டும் என்பது போன்ற தீர்மானங்களுக்கு (பிரஞ்சு comte - comté); எழுத்துரு வடிவ மாற்றதிற்கு (HE READ THE TIMES → He read the times); (உயிரொலிகள் எழுதப்படாத) செமிட்டிக் மொழிகளின் சொல் அணுகலுக்கு (lexical access).

சொற் பொருண்மை மயக்க நீக்கத்தின் சிக்கல் செயற்கை அறிவு முழுமையாக (AI-complete) விளக்கப்படுகிறது; அதாவது பொது அறிவு மற்றும் களஞ்சிய அறிவு இவற்றின் உருப்படுத்தம் போன்ற செயற்கை அறிவின் கடினமான சிக்கல்களுக்கு முதலில் தீர்வுகண்ட பின்புதான் ஒரு சிக்கலை தீர்க்க இயலும். பார்-ஹில்லெலினின் (Bar-Hillel 1960) இயந்திர மொழிப்பெயர்ப்பு பற்றிய மிகவும் அறியப்பட்ட கட்டுரையின் மைய கருத்து பொருண்மை மயக்க நீக்கத்தின் இயற்கையாய் அமையப்பெறும் சிக்கலாகும்; இதில் இவர் *the box in the pen* என்ற வாக்கியத்தில் *pen* என்ற சொல்லின் அர்த்தத்தைத் தானியக்கமாக அறிவதற்கு எந்த வழியும் இல்லை என்று தீர்மானமாகக் கூறுகிறார். பார்-ஹில்லெலின் வாதம் ALPAC அறிக்கைக்கு அடித்தளமிட்டது (ALPAC 1966); இவ்வறிகைதான் 1960-களின் தொடக்கத்தில்

பெரும்பான்மையான இயந்திர மொழிபெயர்ப்பு ஆய்வுகள் கைவிடப்பட்டத்தற்கு நேரடி காரணமாக அமைந்தது.

இருப்பினும் இதே சமயம் அறிவு உருப்படுத்தக் களத்தில் (area of knowledge representation) குறிப்பிடத்தகுந்த முன்னேற்றம், குறிப்பாகப் பொருண்மை வலை அமைப்புகளின் (semantic networks) வரவால் ஏற்பட்டுள்ளது; இவ்வலை அமைப்புகள் சொற் பொருண்மை மயக்க நீக்கத்திற்கு நேரடியாகப் பயன்படுத்தப்படுகின்றது. செயற்கை அறிவு அடிப்படையிலான இயற்கைமொழி புரிந்துகொள்ளுதல் ஆய்வின் (AI-based natural language understanding research) சட்டகத்திலும் பொருளடக்க ஆய்வுக் (content analysis) களங்களிலும் நடையியல் மற்றும் இலக்கிய ஆய்விலும் (stylistics and literary analysis) தகவல் மீட்பிலும் (information retrieval) அடுத்த இருபது ஆண்டுகளாகச் சொற் பொருண்மை மயக்க நீக்க ஆய்வுகள் தொடர்ந்து நடந்தது. கடந்த பத்து ஆண்டுகளாகக் கணினி மொழியியல் களங்களில் நடைபெற்ற பிற செயல்களைப் போன்று, இயந்திரத்தால் படிக்கவியலும் பனுவல்கள் பெரிய அளவில் கிடைப்பதாலும் இதற்கு இணையாகத் தரவுகளின் சீர்மையைப் பற்றிய தகவலைக் கண்டுபிடிப்பதற்கும் பயன்படுத்துவதற்கும் புள்ளியியல் நெறிமுறைகளின் முன்னேற்றத்தாலும் தானியக்கமாகச் சொற் பொருண்மை மயக்க நீக்கம் செய்யும் முயற்சிகள் பெருகியுள்ளன. சொல் வகைப்பாடு பொருண்மை மயக்க நீக்கம் மற்றும் இணை மொழிபெயர்ப்புகளின் வரிசைப்படுத்தம் (alignment of parallel translation) என்ற நெறிமுறைகளுக்கு இணக்கமான பிற சிக்கல்களுக்கும் ஓரளவுக்கு முழுமையாகத் தீர்வு காணப்பட்டதால் சொற் பொருண்மை மயக்க நீக்கம் மைய நிலைக்கு வந்துள்ளது; மேலும் இது இன்று இயற்கைமொழி ஆய்வில் முக்கியமான சிக்கல்களில் ஒன்றாக அடிக்கடி கூறப்படுகிறது. சொற்பொருண்மை மயக்க நீக்க ஆய்வில் ஏற்பட்டுள்ள முன்னேற்றம் காரணமாகவும் சிக்கல்களைத் தீர்க்கும் நெறிமுறைகளின் விரைவான முன்னேற்றத்தாலும் சொற்பொருண்மை மயக்க நீக்க ஆய்வின் நிலையைப்பற்றி மதிப்பிடுவதற்கும் இந்தக் களத்தில் எடுக்கப்பட வேண்டிய அடுத்த நடவடிக்கைகள் பற்றி கருதுவதற்கும் இது பொருத்தமான நிலையாகும். இங்கு சொல் மயக்க நீக்கத்தின் மிக அறியப்பட்ட அணுகுமுறைகள் பற்றியும் எதிர்கால ஆய்வின் திசைகள் பற்றியும் பரந்துபட்ட ஆய்வு செய்யப்படும்.

3.1 சொற்பொருண்மை மயக்க நீக்கத்தின் கருத்தாய்வு

பொதுவாகச் சொற் பொருண்மை மயக்க நீக்கம் ஒரு பனுவலில் (text) அல்லது கருத்தாடலில் (discourse) வரும் சொல்லிற்கு, அச்சொல்லிற்குத் தர இயலும் பிற

அர்த்தங்களிலிருந்து வேறுபடுத்தும் ஒரு வரையறை விளக்கத்தையோ பொருளையோ தருவதாகும். இச்செயல்பாடு இரண்டு நடவடிக்கைகளை உள்ளடக்கும்.

1. கருதப்பட்ட பனுவலுக்கு அல்லது கருத்தடலுக்குப் பொருத்தமாக ஒவ்வொரு சொல்லுக்கும் எல்லா அர்த்தங்களையும் நிறுவுதல்.
2. ஒரு சொல்லின் ஒவ்வொரு நேர்வுக்கும் பொருத்தமான அர்த்தத்தைத் தரும் வழி.

சொல் பொருண்மை மயக்க நீக்கத்தின் அண்மைக்கால ஆய்வு முதல் நடவடிக்கைக்குப் பின்வருவனவற்றை உள்ளடக்கிய முன் வரையறுக்கப்பட்ட அர்த்தங்களைச் சார்ந்திருக்கும்:

- எல்லா அகராதிகளிலும் காணப்படுகின்ற அர்த்தங்களின் பட்டியல்;
- பண்புக்கூறுகள், வகைப்பாடுகள், அல்லது தொடர்புள்ள சொற்கள் இவற்றின் ஒரு குழுமம் [எ.கா. சொற்களஞ்சியத்தில்/பொருட்புல அகராதியில் (thesaurus) ஒருபொருள் பன்மொழிகள் (synonyms)];
- பிற மொழியில் மொழிப்பெயர்ப்புகளை உள்ளடக்கிய மாற்ற அகராதியில் (transfer dictionary) உள்ள ஒரு பதிவு போன்றன.

இருப்பினும் ஒரு அர்த்தத்தின் துல்லியமான வரையறை விளக்கம் ஒரு வாதத்திற்குரிய விசயமாகும். அர்த்தங்களை வரையறை விளக்கம் செய்யும் வேறுபட்ட அணுகுமுறைகள் சொற் பொருண்மை மயக்க ஆய்வின் ஒப்புமைப்படுத்தல் பற்றிய தற்போதைய ஈடுபட்டை எழுப்பியுள்ளது. பொருண்மை வரையறை விளக்கத்தின் சிக்கலின் கடினம் காரணமாகச் சரியான தீர்வு சீக்கிரத்தில் பெறுவது என்பது இல்லை. இருப்பினும் சொற்பொருண்மை மயக்க நீக்கத்தின் தொடக்கக் காலத்திலிருந்தே உருபனியல், தொடரியல் மயக்க நீக்கம் மற்றும் அர்த்த மயக்க நீக்கம் என்பனவற்றின் சிக்கல்களைப் பிரிக்க இயலும் என்பதில் பொதுவான உடன்பாடு இருக்கிறது. அதாவது வேறுப்பட்ட சொல்வகைப்பாடு கொண்ட ஒப்புரு மொழிகளுக்கு (homographs) (எடுத்துக்காட்டாக *play* போன்று பெயராகவும் வினையாகவும் வருவனவற்றிற்கு) உருபனியல், தொடரியல் பொருண்மை மயக்க நீக்கம் அர்த்த மயக்க நீக்கத்தை விளைவிக்கும். எனவே சொல் வகைப்பாட்டு அடையாளப்படுத்திகள் (part-of-speech taggers) உருவாக்கத்திலிருந்தே சொற் பொருண்மை மயக்கம் நீக்கம் செய்யும் ஆய்வுகள் ஒரே தொடரியல் வகப்பாட்டை சார்ந்த ஒப்புரு மொழிகளுக்கு இடையில் அர்த்தங்களை வேறுப்படுத்துவதில் கவனக்குவிப்பு செய்தது.

இரண்டாவது நடவடிக்கை இரு முக்கியமான தகவல் மூலத்தின் மீதான நம்பிக்கை அடிப்படையில் நிறைவேற்றப்படுகிறது:

- இந்த பரந்த அர்த்ததில் சொல்லின் சூழல் மயக்க நீக்கம் செய்யப்பட வேண்டும்: இது சூழல் போன்ற பனுவல் பற்றிய கூடுதல் மொழியியல் தகவலுடன் சொல் தோன்றும் பனுவல் அல்லது கருத்தாலுக்குள் இருக்கும் தகவலை உட்படுத்தும்.
- சொற்களை அர்த்தங்களுடன் தொடர்புபடுத்த உதவும் தரவைத் தரும். சொல் சார்ந்த, களஞ்சியம் சார்ந்த வெளி அறிவுமூலங்கள், மூலவளங்கள், கையால் உருவாக்கப்பட்ட அறிவு மூலங்கள் இவற்றை உட்படுத்தும்.

எல்லாச் சொற்பொருண்மை மயக்கநீக்க ஆய்வுகளும் வெளி அறிவு மூலங்களிலிருந்தான (external knowledge source) தகவல்களுடனோ (அறிவு இயக்கச் சொற்பொருண்மை மயக்கநீக்கம் knowledge driven WSD) அல்லது தரவுத்தொகுதிகளிருந்து எடுக்கப்பட்ட சொல்லின் முந்தைய பொருண்மை மயக்கம் நீக்கப்பட்ட நேர்வுகளின் சூழலைப்பற்றிய தகவல்களுடனோ (தரவால் இயக்கப்பட்ட அல்லது தரவுத்தொகுதி அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் data-driven or corpus-based WSD) பொருண்மை மயக்கநீக்கம் செய்யப்பட வேண்டிய சொல்லின் நேர்வின் சூழலைப் பொருத்துவதை உட்படுத்தும். ஒவ்வொரு சொல் நேர்வுக்கும் ஒரு அர்த்ததை தரவேண்டி, தற்போதைய சூழலுக்கும் தகவலின் இந்த மூலங்களில் ஒன்றிற்கும் இடையில் மிகப் பொருத்தமானதை நிறுவ, சேர்க்கை நெறிமுறைகளின் (association methods) எந்தவகையும் பயன்படுத்தப்படுகின்றது.

3.1.1 இயந்திர மொழிபெயர்ப்பில் தொடக்ககாலச் சொற்பொண்மை மயக்கநீக்கம்

தானியக்க பொண்மை மயக்கநீக்கத்தில் (automated sense disambiguation) முதல் முயற்சிகள் இயந்திர மொழிபெயர்ப்பு சூழலில்தான் செய்யப்பட்டது. வீவர் (Weaver 1949) தன்னுடைய 'memorandum' என்பதில் இயந்திர மொழிபெயர்ப்பில் சொற்பொருண்மை மயக்கநீக்கத்தின் தேவையை விளக்குகிறார்; இந்தத் தலைப்பில் பின்னால் வந்த ஆய்வுகளை உள்ளடக்கிய சொற்பொண்மை மயக்கநீக்க அணுகுமுறையின் அடிப்படையைச் சுருக்கமாகக் கூறுகிறார். எழு மொழிபெயர்ப்பாளர்களுக்குச் சம்பந்தப்பட்ட சொல்லின் இரு பக்கங்களிலும் ஒன்று அல்லது இரண்டு சொற்களைத் தந்து பொருண்மைமயக்கம் உள்ள சொற்களை அவற்றின் உண்மையான உண்மையான சூழலிலும் வேறுபட்ட சூழலிலும் முன்னிலைப்படுத்தி இக்கேள்விக்குக் குறைந்தது பகுதி அளவாவது விடைதர கப்ளானால் (Kaplan 1950) செய்யப்பட்ட மிக அறியப்பட்ட தொடக்காலப் பரிசோதனை முயன்றது. கப்ளான் முழுவாக்கியத்தைத்

தருவதைக் காட்டிலும் சொல்லின் இருபக்கமும் இரு சொற்களைத் தரும் பொருண்மைத் தீர்மானம் (sense resolution) சிறப்பாகச் சொல்லத்தக்கவிதத்தில் நன்மையானதோ போதுமானதோ அல்ல என்று உற்று நோக்கி அறிந்தார். கப்ளானின் ஆய்வுக்குப் பின்னர் பல ஆய்வாளர்களால் இதே கருத்துச்சாயல் கூறப்பட்டது (Masterman 1961, etc.).

ரைஃபுலரின் (Reifler 1955) ஒரு சொல்லுக்கும் சூழலுக்கும் இடையிலான "பொருண்மைசார் எதிர்பாரா பொருத்தங்கள்" ("semantic coincidences") சொற்பொருண்மை மயக்கநீக்கத்தில் நிர்ணியிக்கும் காரணியாக விரைவில் மாறியது. சூழலின் கலவைத் தன்மை மற்றும் குறிப்பாகத் தொடரியல் உறவுகளின் பங்களிப்பும் உணரப்பட்டது. (எடுத்துக்காட்டாக ரைஃபுலர் பின்வருமாறு கூறுகிறார்: இலக்கண அமைப்பும் சொற்பொருண்மை மயக்கநீக்கத்திற்கு உதவ இயலும்; எடுத்துக்காட்டாக *keep* என்ற சொல் அதன் செய்யப்படுபொருள் வினைப்பெயர் (*gerund*) (*He kept eating*), பெயரைடைத் தொடர் (adjectival phrase) (*He kept calm*) அல்லது பெயர்த் தொடர் (noun phrase) (*He kept a record*) என்பதைத் தீர்மானிப்பதால் பொருண்மை மயக்கநீக்கம் செய்யப்பட இயலும்.) தொடக்கத்தில் இயந்திர மொழிபெயர்ப்பின் இலக்கு மிதமானதாக இருந்தது; தொழில் நுட்பப் பனுவல்களின் மொழிபெயர்ப்பின் மீது கவனக்குவிப்பு செய்தது; இந்த எல்லா நேர்வுகளிலும் குறிப்பிட்ட களங்களிலிருந்து/புலங்களிலிருந்து பனுவல்களைக் கையாண்டது. வீவரின் (Weaver,1949) Memorandum பொருண்மை மயக்கநீக்கத்தில் புலத்தின்/களத்தின் பங்களிப்பை விளக்குகிறது; இது பல ஆண்டுகளுக்குப் பின்னர் மீண்டும் வலியுறுத்தப்பட்டது (Gale etal.,1992): மிக எளிதான எடுத்துக்காட்டாகக் கணக்கியலை எடுத்துக்கொண்டால் ஒரு கணக்கியல் கட்டுரையின் பொது சூழலுக்குள் ஒவ்வொரு சொல்லுக்கும் ஒரே ஒரு பொருண்மைதான் இருக்கிறது என்று ஒருவர் ஓரளவுக்குக் கூற இயலும். இந்த உற்றுநோக்கலைப் பின்பற்றி தனிச்சிறப்பு அகராதிகள் (specialized dictionaries) அல்லது நுண்ணிய பொருள் விளக்கச் சொற்கோவைகள் (micro glossaries) இவற்றை உருவாக்குவதில் இயந்திர மொழிபெயர்ப்பின் தொடக்ககாலத்தில் அதிக முயற்சிகள் எடுக்கப்பட்டன. இந்த நுண்ணிய பொருள்விளக்கச் சொற்கோவைகள் கருத்தாடலின் ஒரு குறிப்பிட்ட புலத்தில் பனுவல்களுக்குப் பொருத்தமான/தேவையான தரப்பட்ட சொல்லின் ஒரு பொருளை மட்டுமே கொண்டிருந்தன. எடுத்துக்காட்டாகக் கணிதவியல் புலத்தின் ஒரு நுண்ணிய பொருள்விளக்கச் சொற்கோவை *triangle* (முக்கோணம்) என்பதற்குத் தேவையான வரையறை விளக்கத்தை

மட்டுமே கொண்டிருக்கும்; *triangle* இசைக் கருவியாக இருக்கும் வரையறை விளக்கத்தைத் தராது.

தொடக்கத்திலிருந்தே சொற்பொருண்மை மயக்கநீக்கத்திற்கு அறிவு உருப்படுத்ததின் (knowledge representation) தேவை ஒப்புக்கொள்ளப்பட்டது. மொழிகளின் தர்க்கம் சார்ந்த அமைப்புகளில் மிகக் கூடுதலான ஆய்வு தேவை என்று வீவர் தன்னுடைய Memorandum என்பதில் முடிவுரையாகக் கூறுகின்றார். பல ஆய்வாளர்கள் எந்த மொழியிலும் உள்ள சொற்களை ஒரு பொதுவான பொருண்மையியல் அல்லது கருத்துருசார் உருப்படுத்தத்துடன் பொருத்திச் சொற்பொருண்மை மயக்கநீக்கச் சிக்கலைத் தீர்க்கும் தர்க்கம் மற்றும் கணிதவியல் கொள்கைகளின் அடிப்படையிலான இடைமொழி உபாயத்தை (interlingual device) உருவாக்க முயன்றனர். இந்த முயற்சிகளில் ரிச்சென்ஸ் மற்றும் மாஸ்டர்மேன் (Richens, 1958, Masterman 1961) என்பவர்களின் முயற்சிகள் இறுதியாக 'பொருண்மை வலை அமைப்பு (semantic network) என்ற கருத்துச்சாயலுக்குக் கொண்டுசென்றது. இதைத் தொடர்ந்து முதல் இயந்திர நடைமுறைப்படுத்தப்பட்ட அறிவு அடிப்படை (machine-implemented knowledge base) ராஜஸ்ட் தெசராசிலிருந்து (Rogests thesaurus) உருவாக்கப்பட்டது. மாஸ்டர்மேன் (Masterman 1957) இந்த அறிவு அடிப்படையைச் சொற்பொருண்மை மயக்கநீக்கச் சிக்கலுக்குப் பயன்படுத்தினார்: விர்ஜில் ஜியார்கிக்ஸ் (Virgils Georgics) என்பதை இயந்திரத்தால் மொழிபெயர்க்கும் முயற்சியில் அவர் லத்தின் மொழியின் ஒவ்வொரு சொல்லின் பகுதிக்கும் (word stem) மொழிபெயர்ப்பை லத்தின்-ஆங்கில அகராதியில் பார்த்தார்; பின்னர் ராஜெஸ்டின் சொல்லிலிருந்து-தலைப்பு சொல்லடைவில் (word-to-head index) இச்சொல்லைப்பார்த்தார். இவ்வழியில் ஒவ்வொரு லத்தின் சொற்பகுதியும் ஆங்கில நிகரன்களுடன் தொடர்புடைய ராஜேஸ்ட் தலைப்பு எண்களின் பட்டியலிடன் தொடர்புபடுத்தப்பட்டது. ஒரே வாக்கியத்தில் வரும் சொற்களின் எண்கள் மேலுறல்களுக்காகப் பரிசீலீக்கப்பட்டன. இறுதியாகப் பலதடவை நேர்வு செய்யும் தலைப்பு வகைப்பாடுகளின் கீழ் தோன்றும் ஆங்கிலச் சொற்கள் மொழிபெயர்ப்பிற்காகத் தேர்ந்தெடுக்கப்பட்டன. மாஸ்ட்ர்மேனின் இந்த நெறிமுறை தற்போது நிறைவேற்றப்பட்டுள்ள பல அறிவு அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கத்தில் உள்ளுறைவதுடன் கூறத்தக்க விதத்தில் ஒற்றுமை காட்டுகின்றது.

வீவரின் பனுவலும் கிட்டத்தட்ட 50 ஆண்டுகளுக்குப் பின்னர் தற்போது புழக்கத்தில் உள்ள மொழிக்கான புள்ளியியல் அணுகுமுறையைக் கோடிட்டு காட்டுகிறது என்பது ஆர்வத்துடன்

பார்த்துணர வேண்டியதாகும்: "இந்த அணுகுமுறை சிக்கலின் புள்ளியியல் பண்பு என்று கூறப்படுகின்ற மிக முழுமையான அடிப்படையான செய்தியின் நோக்கை வெளிக்கொணர்கின்றது. […] மேலும் தேவையான முதன்மை நடவடிக்கையாகப் புள்ளியில் பொருண்மையியல் ஆய்வுகள் செய்யப்படவேண்டும் என வலியுறுத்துவது இந்த மெம்மோராண்டத்தின் முக்கியமான நோக்களில் ஒன்றாகும்."

பல ஆய்வாளர்கள் தொடக்ககால இயந்திர மொழிபெயர்ப்பில் இந்த அணுகுமுறையைப் பின்பற்றினார்கள். பனுவல்களிலும் அகராதிகளிலும் பல்பொருண்மையின் அளபின் மதிப்பீடுகள் செய்யப்பட்டன: ரஷ்யப் பனுவல்களில் ஆய்வு செய்த ஹார்பர் (Harper 1957b) ஒரு இயற்பியல் கட்டுரையில் வரும் பல்பொருண்மை சொற்களின் எண்ணிக்கை 30% விழுக்காடு என்றும் மற்றொரு அறிவியல் கட்டுரையில் 43% விழுக்காடு என்றும் நிர்ணயம்செய்தார். அவர் கல்லாஹமின் ரஷ்ய ஆங்கில அகராதி (Callaham's Russian-English Dictionary) ஒவ்வொரு ரஷ்ய சொல்லிற்கும் சராசரியாக 8.6 ஆங்கில நிகரன்கள் தருவதாகக் கண்டுபிடித்தார்; அவற்றில் 5.6 பகுதி-பல்பொருண்மைச் சொற்கள் (quasi-synonyms); இவ்வாறு ஒவ்வொரு ரஷ்யன் சொல்லுக்கும் கிட்டத்தட்ட மூன்று வேறுபட்ட ஆங்கில நிகரன்களைத் தருகின்றது. முதலில் கணினியாக்கம் செய்யப்பட்ட அகராதியில் இரண்டாயிரம் சொற்களில் ஐந்நூறு சொற்கள் பல்பொருண்மைச் சொற்கள் என்று பெல்ஸ்கஜா (Bel'skaja 1957) அறிவித்துள்ளார். பிம்ஸ்லேர் (Pimsleur 1957) மொழிபெயர்ப்பிற்கு ஆழங்களின் மட்டங்கள் (levels of depth for translation) என்ற கருத்துச்சாயலை அறிமுகப்படுத்தினார்: ஒன்றாம் மட்டம் கூடுதல் நிகழும் நிகரனை பயன்படுத்தியது (German schwer = happy); இதன்படி 80% விழுக்காடு சொற்கள் சரியாக மொழிபெயர்க்கப்பட்ட பனுவலை உருவாக்கியது; இரண்டாவது மட்டம் கூடுதலான பொருண்மைகளை வேறுப்படுத்தியது (e.g. schwer = difficult); இதன்படி 90% விழுக்காடு சரியான மொழிப்பெயர்ப்பு உருவாக்கப்பட்டது; இவ்வாறு படிப்படியாக மேம்பட்டுச்சென்றது. கலைச்சொற்கள் வேறுப்பட்டாலும் இது தற்போதைய ஆய்வுகளில் பயன்படுத்தப்படும் அடிப்படை அடையாளப்படுத்தும் (Baseline tagging) கருத்துச்சாயலுடன் ஒற்றுமையுடையது.

இந்தக் கருத்துக்களின் நம்பத்தகுந்த நடைமுறைபடுத்தம் பல ஆண்டுகளுக்குப் பின்னர் செய்யப்பட்டது. இதற்கு முரணாக இயந்திர மொழிபெயர்ப்பு இதே வேளையில் நலிவுறத் தொடங்கியது. மது மற்றும் லைட்டில் (Madhu and Lytle 1965) என்போர் புலம் பொருண்மையைக் கட்டுப்படுத்தும் என்ற உற்றுநோக்கில் ஆய்வு செய்து வேறுப்பட்ட புலங்களின் பனுவல்களுக்கு

அர்த்த நிகழ்வெண்ணை (sense frequency) கணித்தனர். மேலும் ஒருதரப்பட்ட சூழலில் ஒவ்வொரு அர்த்தத்தின் நேர்வு சாத்தியங்களை நிர்ணியிக்க பெய்சன் வாய்ப்பாட்டைப் (Beyesian formula) பயன்படுத்தினார்; இந்தத் தொழில்நுட்பம் மிகப் பிற்காலத்தில் பயன்படுத்தப்பட்ட தொழில்நுட்பத்துடன் ஒற்றுமையுள்ளது; இது ஒரு ஒத்திருக்கிற 90% விழுக்காடு சரியான பொருண்மை மயக்கநீக்க விளைவைத் தந்தது.

சொற்பொருண்மை மயக்கநீக்கத்தின் மீதான இந்த முந்தைய ஆய்வின் சிறப்பான உண்மை, அக்காலத்திலேயே அடிப்படையான சிக்கல்கள் மற்றும் சிக்கல்களுக்கான அணுகுமுறைகள் எதிர்பார்க்கப்பட்டு உருவாக்கப்பட்டதன் அளபாகும். இருப்பினும் மிகப் பெரிய அளவில் மூலவளங்கள் இல்லாத காரணத்தால் இந்தக் கருத்துக்கள் பரிசோதனை செய்யப்படாமல் இருந்தன; மட்டுமல்லாமல் அவை பல ஆண்டுகளுக்குப் பின்னர் வரை பெருமளவில் மறக்கப்பட்டிருந்தன.

3.1.2 செயற்கை அறிவு அடிப்படையிலான நெறிமுறைகள் (AI based methods)

1960-களின் ஆரம்பத்தில் செயற்கை அறிவு நெறிமுறைகள் (AI methods) வளரத் தொடங்கின; அவை மொழி புரிந்துக்கொள்ளல் சிக்கலை கையாளத் தொடங்கின. இதன் விளைவாக முழு மொழியைப் புரிந்துகொள்வதற்கு வேண்டி உருவாக்கப்பட்ட பெரிய ஒழுங்குமுறைகளின் சூழலில் செயற்கை அறிவு ஆய்வில் சொற்பொருண்மை மயக்கநீக்கம் எடுத்துக்காட்டாக நிறைவேற்றப்பட்டது. காலத்திற்கேற்ப இம்மாதிரியான ஒழுங்குமுறைகள் அவர்கள் மாதிரிப்படுத்த முயன்ற மனிதமொழி புரிந்துகொள்ளலின் ஏதாவது கொள்கையின் அடிப்படையாக அமைந்தது. பெரும்பான்மையும் சொற்பொருண்மை மயக்கநீக்கத்திற்குப் பயன்படுத்தப்பட செயல்பாட்டை நிறைவேற்ற வேண்டி தொடரியல் மற்றும் பொருண்மையியல் பற்றிய மிக விளக்கமான அறிவின் பயன்பாட்டை உள்ளடக்கியிருந்தது.

3.1.2.1 குறியீடு சார் நெறிமுறைகள் (symbolic methods)

முன்னர் கூறியபடி பொருண்மை வலையமைப்புகள் (semantic networks) 1950-களில் உருவ்வாக்கப்பட்டன; மற்றும் அவைகள் உடனடியாகச் சொற்பொருண்மைகளை உருப்படுத்தம் செய்யும் சிக்கல்களுக்குப் பயன்படுத்தப்பட்டன. இயந்திர மொழிபெயர்ப்புக் களத்தில் ஆய்வு செய்து வந்த மாஸ்டர்மேன் (Masterman 1961) அடிப்படை மொழி கருத்துருக்கள் (fundamental language concepts) கொண்ட இடைமொழியில் (interlingua) வாக்கியங்களின் உருப்படுத்தத்தை ஆக்கவேண்டி ஒரு பொருண்மை வலையமைப்பைப் (semantic network) பயன்படுத்தினார்;

வலையமைப்பில் மிக நெருக்கமாகத் தொடர்புள்ள கணுக்களின் குழுக்களைப் பிரதிபலிக்கும் உருப்படுத்தங்களைத் தேர்வுசெய்து அர்த்த வேறுப்பாடுகள் உட்கீடையாகச் செய்யப்பட்டன. அவர் 100 மூலக் கருத்துரு வகைகளின் (primitive concept types) (THING, DO) ஒரு குழுமத்தை உருவாக்கினார்; அதன் அடிப்படையில் அவர் 15000 பதிவு கருத்துரு அகராதியை (entry concept dictionary) உருவாக்கினார்; இதில் கருத்துரு வகைகள் உயர் கருத்துருக்களிருந்து துணைக்கருத்துருக்களுக்குப் பண்புக்கூறுகளின் மரபுரிமையால் ஒரு பின்னல்வேலைப்பாட்டில் ஒழுங்குபடுத்தப்பட்டுள்ளன. இதன் மீதும் ரிச்சென்சின் (1958) பொருண்மை வலையமைப்புகள் ஆய்வின் மீதும் குல்லியன் (Qullian 1961, 1962 a, b, 1967, 1968, 1969) சொற்களுக்கும் (‘tokens’) கருத்துருக்களுக்கும் (“types”) இடையில் தொடர்புகளை உள்ளடக்கிய ஒரு வலையமைப்பை உருவாக்கினார்; இதில் தொடர்புகள் பல்வேறுப்பட்ட பொருண்மை உறவுகளாலோ அல்லது சொற்களுக்கு இடையிலான தொடர்புகளைக் காட்டுவதாகவோ புலப்படுத்தம் செய்யப்பட்டுள்ளது. வலையமைப்புகள் அகராதிவரையறை விளக்கங்களிலிருந்து தொடங்கி உருவாக்கப்பட்டுள்ளது; இருப்பினும் கை குறியனாக்கம் செய்யப்பட்ட (hand-coded) மனித அறிவால் மேம்படுத்தப்பட்டுள்ளது. வலையமைப்பில் இரு சொற்கள் தரப்படும்போது குல்லியனின் வழியமைப்பு, குறிகடத்தலின் (marker passing) வழி ஒவ்வொரு உள்ளீட்டுச் சொல்லிலிருந்தும் தொடங்கும் தொடர்புகளின் ஒரு வழியிலூடே கருத்துருக் கணுக்களில் படிப்படியான ஊக்கத்தைப் போலச்செய்யும். இரண்டு உள்ளீட்டுச் சொற்களுக்கு இடையில் காணப்படும் மிக நேரடியான வழியில் உட்படுத்தப்படக்கூடும் தரப்பட்ட உள்ளீட்டுச் சொல்லுடன் ஒரே ஒரு கருத்துரு தொடர்புபடுத்தப்படுவதால் சொற்பொருண்மை மயக்கநீக்கம் நிறைவேற்றப்படுகின்றது. குல்லியனின் ஆய்வு சொற்பொருண்மை மயக்க நீக்கத்திற்கு அகராதி அடிப்படையிலான அணுகுமுறைகளைப் பின்னர்கூறியது.

இதைத் தொடரும் செயற்கையறிவு அடிப்படையிலான அணுகுமுறைகள் சொற்களைப் பற்றியும் தனி வாக்கியங்களில் பிற சொற்களுடன் அவற்றின் பங்களிப்புகளையும் உறவுகளையும் பற்றியும் தகவல்களைக் கொண்டிருக்கும் சட்டகங்களின் பயனைச் சாதகமாகப் பயன்படுத்திக்கொண்டுள்ளது. எடுத்துக்காட்டாக ஹெய்ஸ் (Hayes 1976, 1977, 1978) பொருண்மை வலையமைப்பு மற்றும் வேற்றுமைச் சட்டகங்களின் ஒருங்கிணைப்பைப் பயன்படுத்துகிறார். இந்த வலையமைப்பு (network) பெயர் அர்த்தங்களை (noun sense) உருப்படுத்தம் செய்யும் கணுக்களையும் வினை அர்த்தங்களால் (verb sense) உருப்படுத்தம்

செய்யப்படும் தொடர்புகளையும் (links) கொண்டிருக்கும்; வேற்றுமை சட்டகங்கள் (case frames) வலையமைப்பின் மீது 'இரு' (IS-A) உறவையும் 'சினை' (PART-OF) உறவையும் வற்புறுத்தும். குல்லியன் ஒழுங்குமுறையில் உள்ளது போல் சொற்களுக்கு இடையில் தொடர்புகளின் சங்கிலிகளைக் கண்டுபிடிக்க வலையமைப்பு ஊடுருவப்படும். ஹேயஸின் ஆய்வு இந்த அணுகுமுறையைப் பயன்படுத்தி ஒப்புருச்சொற்களை ஓரளவுக்கு துல்லியமாகப் பொருண்மை மயக்கநீக்கம் செய்யலாம் எனக்காட்டுகின்றது; இருப்பினும் பல்பொருண்மையின் (பல்பொருளொருமொழியத்தின்) பிற வகைகளுக்கு குறைந்த அளவுதான் வெற்றித்தரும். ஹிர்ஸ்டும் (Hirst 1987) சட்டகங்களின் வலையமைப்பைத்தான் பயன்படுத்துகிறார். இவரும் குல்லியனைப் பின்பற்றிச் சூழலில் அர்த்தங்களுக்குள் பொருத்தமானதைத் தெரிவுசெய்யச் சொற்களின் அர்த்தங்களின் சட்டகங்களுக்கு இடையில் சேர்க்கையின் குறைந்த நீள வழிகளைக் (minimum length paths of association) கண்டுபிடிக்க வேண்டி குறியீடு கடத்தலைப் (marker passing) பயன்படுத்துகிறார். அவர் இருதுருவ அமைப்புச் சொற்கள் (polaroid words) என்பதை அறிமுகப்படுத்துகிறார்; இந்த இயக்கமுறை சட்டக வலையமைபில் காணப்படும் பொருண்மை உறவுகளுடன் பகுப்பானால் தரப்படும் தொடர்பியல் சான்றின் அடிப்படையில் பொருத்தமில்லாத அர்த்தங்களை நீக்கும். இறுதியாக ஒரு அர்த்தம் இருக்கும்; இருப்பினும் ஹிஸ்ட் உருவகமாக, ஆகுபெயராக அல்லது அறியாத அர்த்தங்களில் வாக்கியங்களில் பயன்படுத்தப்படும் சில சொற்களின் நேர்வுகளில் இருதுருவ அமைப்புகள் பெரும்பாலும் எல்லாச் சாத்தியமான அர்த்தங்களையும் நீக்கிவிடுகின்றன; இதனால் உதவாது போகும் என்கின்றார்.

மாஸ்டர்மானின் தொடக்க நிலைகளைப் (primitives) பயன்படுத்தும் வில்க்சின் விருப்பத்தேர்வுப் பொருண்மையியல் (preference semantics) (Wilks 1968, 1969, 1973, 1975-a-d) இயற்கைமொழி புரிதலுக்கு முக்கியமாக வேற்றுமை அடிப்படையிலான அணுகுமுறையாகும் (Case-based approach); இது சொற்பொருண்மை மயக்கநீக்கச் சிக்கலைக் கையாள வேண்டிச் சிறப்பாகத் திட்டமிடப்பட்டுள்ள ஒழுங்குமுறைகளில் ஒன்றாகும். விருப்பத்தேர்வுப் பொருண்மையியல் ஒரு வாக்கியத்தில் சொற்களின் ஒன்றிணைப்புகளுக்குத் தேர்வுக்கட்டுப்பாடுகளைக் குறிப்பிட்டுக் கூறுகின்றது; விருப்பத்தேர்வு செய்யப்பட்ட கட்டுப்பாடுகள் தோன்றாவிட்டால் இதைத் தளத்த இயலும்; இதனால் குறிப்பாக உருவகங்களைக் கையாள இயலும் (எ.கா. எனது கார் பெட்ரோல் குடிக்கின்றது/*My car drinks gasoline*; இங்கு குடி/*drink* மீதான தேர்வுக்கட்டுப்பாடு ஒரு விலங்கின எழுவாயை விருப்பத்தேர்வு செய்யும்;

ஆனால் ஒரு விலங்கினமல்லா எழுவாயை அனுமதிக்கும்). பொகுரெவ் (Boguraev 1979) விருப்பத்தேர்வு பொருண்மையியல் பல்பொருள் ஒருமொழியத்தன்மை உள்ள வினைகளைக் கையாள ஏற்புடையதல்ல என்று காட்டுகிறார்; அவர் தேர்வுக்கட்டுப்பாடுகள், விருப்பத்தேர்வுகள், வேற்றுமைச் சட்டகங்கள் போன்றவை உள்ளடங்கிய எடுத்துகாட்டின் ஒருங்கிணைப்பைப் பயன்படுத்தி வில்க்சின் நெறிமுறையை மேம்படுத்த முயல்கின்றார். அவர் தரப்பட்ட அர்த்த ஒதுக்கீட்டின் பொருண்மை இயைபைப் பற்றியத் தீர்மானங்களைச் சாத்தியமாக்கச் சொற்பொருண்மை மயக்கநீக்கத்தை அமைப்புப் பொருண்மை மயக்கநீக்கத்துடன் ஒன்றுசேர்க்கின்றார். இக்காலகட்டத்தின் பிற ஒழுங்குமுறைகளைப் போல் இந்த ஒழுங்குமுறைகளும் வாக்கிய அடிப்படையிலானது; இது தலைப்பு மற்றும் பொருட்புலத் தகவல் போன்ற கருத்தாடலின் பிற நிலைகளில் உள்ள நிகழ்வுகளுக்கு விளக்கமளிக்கவில்லை. சில வகை பொருண்மை மயக்கநீக்கங்களை நிறைவேற்றுவதற்குக் கடினமாக இருந்தது அல்லது இயலாமல் இருந்தது என்பது இதன் விளைவாகும்.

கணிசமான அர்த்த வேறுபாட்டுப் பகுதியைக் (sense discrimination component) கொண்டிருக்கின்ற மொழி புரிந்துகொள்ளுதலுக்கு ஒரு வேறுபாடான அணுகுமுறை, சொல் திறன்மிகு பகுப்பான் (Word Expert Parser) (Small 1980, 1983; Small and Riger, 1982; Adriaens, 1986, 1987, 1989; Adriaens and Small 1988) ஆகும். இந்த அணுகுமுறை மொழியைப் பற்றிய மனித அறிவு, விதிகளுக்குப் பதிலாகச் சொற்களைப் பற்றிய அறிவாக முதன்மையாக ஒழுங்குபடுத்தப்பட்டுள்ளது என்ற மரபுக் கோட்பாட்டிலிருந்து ஆகப்பட்டது. அவர்களின் ஒழுங்குமுறை மனித மொழிப் புரிதல் செயற்பாங்கு என்று அவர்கள் நினைப்பதை மாதிரிப்படுத்துகின்றது: தொடரியலும் பொருண்மையியலும் கேள்விக்குரிய சூழலில் அவை ஒவ்வொன்றின் பங்களிப்பைத் தீர்மானிப்பதால் பொருண்மையியல் மற்றும் தொடரியல் பற்றிய சொல் வல்லுனர்களுக்கிடையே (word experts) தகவல் கருத்துப் பரிமாற்றத்தின் ஒருங்கிணைவு. ஒவ்வொரு வல்லுனரும் சொல்லின் எல்லா அர்த்தங்களுக்கும் வேறுபடுத்தும் வலையைக் (discrimination net) கொண்டிருக்கின்றது; இது சூழலாலும் பிற வல்லுனர்களாலும் தரப்பட்ட தகவலின் அடிப்படையில் கடக்கப்படுகின்றது; இறுதியாக வாக்கியத்தின் பொருண்மை உருப்படுத்தத்திற்குச் சேர்க்கப்படும் தனித்தன்மையான அர்த்தத்தை அடைகின்றது. இந்த ஒழுங்குமுறையின் மிக அறியப்பட்ட குறைபாடு என்னவென்றால் இலக்கை அடைவதற்குச் சொல் வல்லுனர்கள் அதிக அளவிலும் கலவைத்தன்மையாகவும் இருப்பது அவசியமாகும்; இது பொருண்மை மயக்கநீக்கத்தைவிட மிக அதிகமாகும்.

டல்க்ரெனின் (Dahlgren 1988) மொழிப் புரிதல் ஒழுங்குமுறை பல வகைப்பட்ட தகவல்களைப் பயன்படுத்தும் பொருண்மை மயக்கநீக்கப் பகுதியை உட்படுத்தும்: நிலையான தொடர்கள், தொடரியல் தகவல்கள் (முதன்மையாகத் தேர்வுக்கட்டுப்பாடுகள்) மற்றும் பொது அறிவு காரணநியாயம் (commonsense reasoning). காரணநியாயத் தொகுதி, அது கணினி அடிப்படையில் தீவிரமானது என்பதால் பிற இரு நெறிமுறைகள் முடிவைத் தரத் தவறும் போது மட்டுமே வேண்டப்படும்/அழைக்கப்படும். அவரது மூல அனுமான பொருண்மை மயக்கநீக்கத்தின் பெரும்பான்மையும் பத்தித் தலைப்பு அடிப்படையில் எய்த இயலும் என்பதாக இருந்தாலும், உண்மையில் பொருண்மை மயக்கத்தின் அரைப் பகுதி நிலையான தொடர் மற்றும் தொடரியல் தகவலைப் பயன்படுத்தி எய்தப்பட்டது என்றும் அடுத்த அரைப்பகுதி பொது அறிவு காரணநியாத்தைப் பயன்படுத்தி எய்தப்பட்டது என்றும் கண்டுகொண்டார். காரணநியாயம் சூழலில் சொற்களுக்குப் பொதுவான முன்னோர்களைக் கண்டுபிடிக்க மெய்பொருள் மூல ஆய்வை (மூலப்பொருண்மை ஆய்வு) (ontology) துழாவுவதை உட்படுத்தும்; மூலப்பொருண்மை ஆய்வில் பொதுவான முன்னோரை உட்படுத்தும் மூலப்பொருண்மை ஒற்றுமை ஒரு திறனுள்ள பொருண்மைநீக்கி என்பதை உறுதிசெய்வதால்/கண்டுபிடிப்பதால் அவரது ஆய்வு ரெஸ்னிக்கின் (Resnik 1993a,b; 1995a) முடிவுகளை எதிர்பார்க்கின்றது. வினை தேர்வுக்கட்டுப்பாடுகள் பெயர்களுக்கு பொருண்மை மயக்கநீக்கத் தகவலின் முக்கியமான மூலபொருளாகும் என்று அவர் கண்டறிந்துள்ளார்; இந்த மற்றொரு முடிவு தொடர்ந்து பரிசோதிக்கப்பட்டுக் குறித்துக்கொள்ளப்பட்டது.

3.1.2.2. தொடர்புபடுத்துபவர் நெறிமுறைகள் (Connectionist methods)

1960களிலும் 1970களிலும் உளமொழியியல் ஆய்வு பொருண்மை முதன்மையாக்கம் (semantic priming) மனிதர்களால் செய்யப்படும் பொருண்மை மயக்கநீக்கத்தில் பங்களிப்பு செய்கின்றது என்று நிறுவியது. பொருண்மை முதன்மையாகச் செயல்பாட்டில் சில கருத்துருக்களை அறிமுகப்படுத்துவது பொருண்மை அடிப்படையில் உறவுள்ள தொடர்ந்து அறிமுகப்படுத்தப்படும் கருத்துருக்களின் செயற்பாங்கைப் பாதிக்கவும் வசதிசெய்யவும் செய்யும் (பார்க்க Meyer and Schvaneveldt, 1971). இந்தக் கருத்து பரவும் செயலாக்க மாதிரிகளில் (spreading activation models) (Collins and Loftus 1975; Anderson, 1976, 1983) உருப்படுத்தம் செய்யப்பட்டுள்ளது; இங்கு பொருண்மை வலையமைப்பில் (semantic network) உள்ள கருத்துருக்கள் பயன்பாட்டால் செயலாக்கப்படுத்தப்பட்டுள்ளது மற்றும் செயலாக்கம் தொடர்புபடுத்தப்பட்ட கணுக்களுக்குப் பரவும். செயலாக்கம் பரவும் போது வலுவிழக்கும்;

ஆனால் சில கணுக்கள் பல மூலங்களிலிருந்து செயலூக்கத்தைப் பெறக்கூடும் மற்றும் படிப்படியாகப் பலப்படுத்தப்படும். மெக்ளிலாண்ட் மற்றும் ருமெல்ஹார்ட் (McClelland and Rumelhart 1981) கணுக்களுக்கிடையில் தடை (inhibition) கருத்துச்சாயலை அறிமுகப்படுத்தி மாதிரியில் சேர்த்தனர்; இதில் ஒரு கணுவின் செயலூக்கம் அவற்றின் சில அடுத்துவருபவைகளைச் செயலூக்கப்படுத்துவதற்குப் பதிலாக அடக்கும் (பார்க்க Feldman and Ballard 1992). சொற்பொருண்மை மயக்கநீக்கத்திற்கு இதைப் பயன்படுத்தும் போது இந்த அணுகுமுறை ஒரு கருத்துருவுக்குப் பொருந்துகிற ஒரு கணுவை செயலூக்கப்படுத்துவது, எடுத்துக்காட்டாக THROW என்ற கருத்துருவைச் செயலூக்கப்படுத்தும் போது பந்து/ball என்பதன் "பௌதிகப் பொருள்" அர்த்தம் செயலூக்கப்படுத்தப்படும்; இதன் செயலூக்கப்படுத்தம் திருப்பவும் "சமூக நிகழ்வு" போன்ற பந்து/ball என்பதன் பிற அர்த்தங்களின் செயலூக்கப்படுத்தத்தைத் தடைசெய்யும்.

மேலே விளக்கப்பட்ட குல்லியனின் (Quillian) பொருண்மை வலையமைப்பு, சொற்பொருண்மை மயக்கநீக்கத்திற்குப் பயன்படுத்தப்பட்ட பரவும் செயலூக்க வலையமைப்பின் (spreading activation network) மிகத் தொடக்ககால நடைமுறைபடுத்தம் ஆகும். இதனுடன் ஒற்றுமையுள்ள மாதிரி காட்ரல் மாற்றும் ஸ்மால் (Cotttrell and Small 1983) என்பவர்களால் நடைமுறைப் படுத்தப்பட்டுள்ளது. இந்த இரு மாதிரிகளிலும் வலையமைப்புகளில் உள்ள ஒவ்வொரு கணுவும் ஒரு குறிப்பிட்ட சொல்லையோ கருத்துருவையோ உருப்படுத்தம் செய்யும். வால்ட்ஸ் மாற்றும் போலாக் (Walt and Pollack 1985) மற்றும் புக்மான் (Bookman 1987) என்போர் அடிப்படை பொருண்மை வேறுபாடுகளுடன் பொருந்தும் பொருண்மை "நுண்பண்புக்கூறுகளின்" ("microfeatures") குழுமங்கள் (விலங்கினம் / விலங்கினம் அல்லாதன, உண்ணக்கூடியது / உண்ணவியலாதது, அபாயகரமானது / பாதுகாப்பானது போன்றன), நிகழ்வுகளின் சிறப்பியல்பான கால அளவு (செகண்ட், மினிட், மணி, நாள் போன்றன), இடங்கள் (பெருநகர், நாடு, கண்டம் போன்றன) மற்றும் இது போன்ற வேறுபாடுகளைத் தங்கள் வலையமைப்பில் கைக்குறியாக்கம் செய்துள்ளனர். வால்ட்ஸ் மற்றும் போலாக்கில் (Walt and Pollack 1985) தொடர்ந்து உள்ளீடு செய்யப்பட்ட சொல்லைப் பொருண்மை மயக்கநீக்கம் செய்வதற்குச் சூழலைச் செயலூக்கப்படுத்த பயன்படுத்துபவரால் நுண்பண்புக்கூறுகளின் குழுமங்கள் கையால் முதன்மைப் படுத்தப்படவேண்டும்; ஆனால் புக்மான் (Bookman 1987) முன்வரும் பனுவலால் நுண்பண்புக்கூறுகள் தானியக்கமாகச் செயலூக்கப்படுத்தப்படும், இவ்வாறு தற்காலிகச் சூழல் நினைவகமாகச் செயல்படும் இயக்கச் செயற்பாங்கை

விளக்குகின்றார். இருப்பினும் இந்த வட்டார மாதிரிகளைக் (local models) காரணகாரியமாக உருவாக்க இயலும் என்றாலும் வினியோகிக்கப்பட்ட மாதிரிகள் பொருண்மை மயக்கநீக்கம் செய்யப்பட்ட எடுத்துக்காட்டுளைப் பயன்படுத்தும் ஒரு கற்றல் கட்டத்தை வேண்டும்; இது அவற்றின் நடைமுறையை எல்லைப்படுத்தும்

செயற்கை அடிப்படையிலான ஒழுங்கமைப்புகளுக்குத் தேவையான அறிவு மூலங்களைக் கையால் உருவாக்குவதன் கடினம், மொழியின் மிகக் குறைந்த அளவைக் கையாளும் "பொம்மை" நிறைவேற்றல்களுக்கு அவற்றை எல்லைப்படுத்துகின்றன. இதன் விளைவாக அது போன்ற ஒழுங்குமுறைகளில் உள்ளுறையும் பொருண்மை மயக்கநீக்கச் செயன்மைகள் எல்லைக்குட்பட்ட சூழலில் (பெரும்பான்மையும் ஒரு தனி வாக்கியம்) மிகச்சிறிய பரிசோதனை குழுமத்தில் மட்டுமே மிக வழக்கமாக பரிசோதிக்கப்பட்டுள்ளன. குறைந்த வெளிப்படையான காரணங்களுக்காகப் பல செயற்கை அறிவு அடிப்படையிலான பொருண்மை மயக்கநீக்க முடிவுகள் மிக அதிகமான பொருண்மை மயக்கம் உள்ள சொற்களையும் நுட்பமான பொருண்மை வேறுபாடுகளையும் (எ.கா. ask, idea, move, use, work, etc.) உட்படுத்தும்; வாக்கியங்களைப் பரிசோதிப்பது சாத்தியமில்லை (The astronomer married the star); இது முடிவுகளைப் பொருண்மை வேறுபாடுளை வேறுபடுத்தும் அண்மையில் அறியப்பட்ட கடினங்களின் அடிப்படையில் மதிப்பீடு செய்ய எளிமையற்றதாக மாற்றுகின்றது.

3.1.3 அறிவு அடிப்படையிலான நெறிமுறைகள் (Knowledge based methods)

70களிலும் 80களிலும் செயற்கை அறிவு அடிப்படையிலான ஆய்வுப்பணிகள் கோட்பாடு அடிப்படையில் ஆர்வமூடுபவை; ஆனால் மிக எல்லைக்குட்பட்ட பொருட்புலங்கள் தவிர வேறு எந்தப் பொருட்புலத்திலும் மொழிப் புரிதலுக்கு நடைமுறைக்குரியது அல்ல. சொற்பொருண்மை மயக்கநீக்க ஆய்வைப் பொதுமையாக்குவதற்கான குறிப்பிடத்தகுந்த தடை, சொற்பொருண்மை மயக்கநீக்கத்திற்கு தேவையான அறிவின் மிகப்பெரிய அளவைக் கையால் செய்வதன் கடினமும் விலையும் ஆகும்: "அறிவுப் பேறு முட்டுக்கட்டை" ("knowledge acquisition bottleneck") என்று அழைக்கப்படுவது ஆகும் (Gale et al, 1993). 1980இல் அகராதிகள், பொருட்புல அகராதிகள், தரவுத்தொகுதிகள் போன்ற பெரிய அலவிலான சொல்சார் மூலவலங்கள் பரந்த அளவில் கிடைக்கத் தொடங்குகையில் சொற்பொருண்மை மயக்கநீக்கத்தின் ஆய்வுப்பணிகள் ஒரு திருப்புமுனையை அடைந்தது. இந்த மூலங்களிலிருந்து தானியக்கமாக அறிவைப் பிரித்தெடுக்க முயற்சிகள் தொடங்கப்பட்டன; அண்மைக் காலத்தில் கையால் பெரிய அளவிலான அறிவு அடிப்படைகளை (knowledge base) உருவாக்குவதற்கு முயற்சிகள் எடுக்கப்பட்டுள்ளன. இதற்கு இணையாக இக்கால கட்டத்தில் மொழியியல் கோட்பாடுகள் அடிப்படையிலான

நெறிமுறைகளிலிருந்து விலகி அனுபவ நெறிமுறைகள் அடிப்படைகளுக்கு மாற்றம் நிகழ்ந்தது; மட்டுமன்றி முழு ஒழுங்குமுறையையும் உருவாக்குவதற்கு முக்கியத்துவம் தருவது குறைந்து அதற்குப் பதிலாகச் சொற்பொருண்மை மயக்கநீக்கம் போன்ற "இடைநிலை" ஆய்வுப்பணிகளைச் செய்வதற்கு முக்கியத்துவம் அளிக்கப்பட்டது.

3.1.3.1 இயந்திரம் படிக்கவியலும் அகராதிகள்

அம்ஸ்லர் (Amsler 1980) மற்றும் மைக்கேல் (Michiels 1982) என்பவர்களின் ஆய்வேடுகளைப் பின்பற்றி மொழி ஆய்வுப் பணிகளுக்குப் பிரபலமான அறிவு மூலமாக/ஆதாரமாக இயந்திரம் படிக்கவியலும் அகராதிகள் (Machine Redable Dictionaries (MRDs) மாறின. 1980களின் காலகட்டத்தில் செயல்பாட்டின் முக்கிய களம், இயந்திரம் படிக்கவியலும் அகராதிகளிலிருந்து சொல்சார் மற்றும் பொருண்மையியல் அறிவைத் தானியக்கமாகப் பிரித்தெடுப்பதற்கான முயற்சிகளை உட்படுத்தியது (Michiels et al 1980; Calzolari 1984; Chodorow et al 1985; Markowizt et al 1986; Byrd et al 1987; Nakamura and Nagao 1988; Klavans et al 1990; Wilks et al 1990 etc.). இந்த ஆய்வுப்பணி பொருண்மையியல் ஆய்வுகளுக்கு சிறப்பாகப் பங்களிப்பு செய்தது; ஆனால் ஆரம்ப நோக்கமான 'பெரிய அளவிலான அறிவு அடிப்படைகளின் தானியக்கப் பிரித்தெடுப்பு' முழுவதும் எய்தப்பெறவில்லை: தற்போதுதான் பரவலாகக் கிடைக்கின்ற பெரிய அளவிலான அறிவு அடிப்படை (சொல் வலை (wordNet)) கையால் உருவாக்கப்பட்டது.

குறைபாடுகள் இருந்தாலும் இயந்திரத்தால் படிக்கவியலும் அகராதிகள் சொற்களின் அர்த்தங்களைப் பற்றிய தகவல்களின் தயார்நிலை ஆதாரத்தைத் தருகின்றது; எனவே சொற்பொருண்மை மயக்கநீக்கத்தின் முக்கிய ஆதாரப் பொருளாக மாறியுள்ளது. பயன்படுத்தப்பட்டுள்ள நெறிமுறைகள் ஒரு தரப்பட்ட அகராதியின் சீரின்மையை குறைப்பதற்கு அல்லது நீக்குவதற்குப் போதுமான அளவு பலமான நெறிமுறைகளுடன் நேரடியாக அகராதி வரையறை விளக்கங்களின் பனுவல்களைப் பயன்படுத்தி முன்னர் கூறப்பட்ட சிக்கல்களை விலக்க முயற்சிகள் எடுத்துள்ளன. பன்னிலையில் சேர்ந்துவரும் சொற்களுக்குக் கூடுதல் சாத்தியமான அர்த்தத்தைத் தருவது, தேர்ந்தெடுக்கப்பட்ட அர்த்தங்களுக்குள் உறவுத்தன்மையை அதிகப்படுத்தும் ஒன்று என்ற கருத்துச்சாயலை எல்லா நெறிமுறைகளும் சார்ந்திருக்கிறது.

லெஸ்க் (Lesk 1986) ஒரு அகராதியில் ஒவ்வொரு அர்த்தத்துடனும் தொடர்புடைய ஒரு அறிவு அடிப்படையை உருவாக்கினார்; ஒரு ஒப்பம் (signature) அந்த அர்த்ததின் வரையறை

விளக்கத்தில் தோன்றும் சொற்களின் பட்டியலைக் கொண்டது. ஒரு சொல்லின் சூழலில் அடுத்துவரும் சொற்களின் ஒப்பங்களுடன் மிகக்கூடுதல் எண்ணிக்கையில் ஒப்பங்களின் மேலுறல்களைக் கொண்டிருக்கும் இலக்குச் சொற்களின் அர்த்தத்தைத் தேர்ந்தெடுப்பதால் சொற்பொருண்மை மயக்கநீக்கம் எய்தப்படுகின்றது. ஒரு எடுத்துக்காட்டான கற்பவர் அகராதியில் காணப்படுவது போன்ற அர்த்த வேறுபாடுகளின் ஒத்தறி அடிப்படையில் நல்ல குழுமத்தைப் பயன்படுத்தி இந்த நெறிமுறை 50-70% சரியான பொருண்மை மயக்கநீக்கத்தை எய்துகின்றது. லெஸ்கின் நெறிமுறை ஒவ்வொரு வரையறை விளக்கத்தின் சரியான சொல் பயன்பாட்டிற்கு மிகவும் நுண்ணுர்வுள்ளது: தரப்பட்ட ஒரு சொல்லின் இருப்பு அல்லது இல்லாமை தீவிரமாக முடிவை மாற்றும். இருப்பினும் லெஸ்கின் நெறிமுறை பல இயந்திரத்தால் படிக்க இயலும் அகராதி அடிப்படையிலான பொருண்மை மயக்கநீக்க ஆய்வுப்பணிகளுக்கு அடிப்படையாக அமைந்தது.

3.1.3.2 பொருட்புல அகராதிகள் (சொற்களஞ்சியங்கள்)

பொருட்புல அகராதிகள் (thesaurus) சொற்களுக்கு இடையில் உள்ள உறவுகளைப் பற்றி, குறிப்பாக ஒருபொருள் பன்மொழியத்தைப் பற்றித் தகவலைத் தருகின்றது. 1950களில் இயந்திரத்தால் படிக்கவியலும் வடிவில் மாற்றப்பட்ட ரோஜெஸ்டின் அனைந்துலகப் பொருட்புல அகராதி (Roget's Intenational Thesarus) இயந்திர மொழிபெயர்ப்பு (machine translation) (Masteman, 1957), தகவல் மீட்பு (information retrieval) (Sparck-Jones 1964, 1986) மற்றும் பொருளடக்க ஆய்வு (content analysis) (Sedelow and Sedelow 1969, 1986, 1992) உள்ளடக்கிய பல்வேறு பயன்பாடுகளில் பயன்படுத்தப்பட்டது; இது ஏறுமுகமாக சீராக்கப்பட்ட எட்டு மட்டங்கள் வரை கொண்ட வெளிப்படையான கருத்துருப் படிநிலை அமைப்பைத் தருகின்றது. எடுத்துக்காட்டாகப் பொருட்புல அகராதியின் வேறுபட்ட வகைபாடுகளின் கீழ் வரும் ஒரே சொல்லின் ஒவ்வொரு நேர்வும் அச்சொல்லின் வேறுபட்ட அர்த்தங்களை உருப்படுத்தம் செய்யும்; வகைப்பாடுகள் ஓரளவுக்குச் சொல்லின் அர்த்தங்களுடன் பொருத்தமுறும் (Yarowsky 1992). ஒரே வகைப்பாட்டில் உள்ள ஒரு குழுமச் சொற்கள் பொருண்மை அடிப்படையில் தொடர்புள்ளனவை ஆகும்.

தொடக்கத்தில் அறியப்பட்ட ரோஜஸ்டின் சொற்பொருண்மை மயக்கநீக்கப் பயன்பாடு முன்னர் விளக்கப்பட்ட மாஸ்டர்மானின் ஆய்வுப்பணியாகும் (Masterman 1957). பல ஆண்டுகளுக்குப் பின்னர் பொருட்புல அகராதியிலிருந்து ஆக்கப்பட்ட எ-சங்கலிகளால் (e-chains) உருவாக்கப்பட்ட பொருண்மைக் கொத்துகளைப் (semantic clusters) பரிசோதித்து வினைகளின்

அர்த்தங்களை வேறுபடுத்த பாட்ரிக் (Patrick 1985) ரோஜஸ்டைப் பயன்படுத்தினார் (Bryan 1973, 1974; Sedelow and Sedelow 1986). அவர் பொருட்புல அகராதியில் பொருண்மை அடிப்படையில் மிக நெருக்கமாக உறவுகொண்டுள்ள கீழ்-மட்ட அரைப்புள்ளி குழுமங்களில் (low-level semicolon groups) சொற் குழுமங்களைக் கொண்ட "சொல்-வலுவான அக்கம்பக்கங்கள்" (word-strong neighbourhood) மற்றும் சங்கலிகள் வழி அக்குழுமங்களுடன் இணைக்கப்பட்டுள்ள சொற்கள் இவற்றைப் பயன்படுத்துகிறார்.

யரோவ்ஸ்கி (Yarovsky 1992) ரோஜஸ்டில் பொதுவான வகைப்பாடுகளில் உள்ள சொற்களில் தொடங்கி சொற்களின் வகுப்புகளை ஆக்குகின்றார். வகைப்பாட்டின் ஒவ்வொரு சொல்லின் 100-சொல் சூழலும் தரவுத்தொகுதியிலிருந்து பிரித்தெடுக்கப்பட்டுள்ளன; பரஸ்பர-தகவல்-போன்ற புள்ளியியல் அந்த வகைப்பாட்டு உறுப்பினருடன் சேர்த்துவரும் கூடுதல் சாத்தியமான சொற்களைக் கண்டுபிடிக்கப் பயன்படுத்தப்படுகின்றது. விளையும் வகுப்புகள் பல்பொருள் ஒருமொழியத்தின் புதிய நேர்வுகளைப் பொருண்மை மயக்கம் நீக்கம் செய்யப் பயன்படுத்தப்படுகின்றது: பல்பொருள் ஒருமொழிய நேர்வின் 100-சொல் சூழல் பல்வேறு வகுப்புகளில் உள்ள சொற்களுக்குப் பரிசோதிக்கப்பட்டது; பல்பொருள் ஒருமொழியின் கூடுதல் சாத்தியமான வகுப்பைத் தீர்மானிக்கப் பெய்ஸ் விதி (Bayes' Rule) பயன்படுத்தப்பட்டது. யரோவ்ஸ்கியால் வகுப்பு ஒரு சொல்லின் குறிப்பிட்ட அர்த்தத்தை உருப்படுத்தம் செய்வதாகக் கருதப்பட்டதால் ஒரு வகுப்புக்கு ஒத்துக்குவது அர்த்தத்தைக் கண்டுபிடித்தது. அவர் சராசரி 3-வழி அர்த்த வேறுபாட்டின் மீது 92% துல்லியத்தை அறிவிக்கிறார். யரோவ்ஸ்கி அவருடைய நெறிமுறை தலைப்புத் தகவலைப் பிரித்தெடுக்க நல்லது என்கிறார்; இது பெயர்களின் பொருண்மை மயக்கநீக்கத்திற்கும் அதிக வெற்றியைத் தரக்கூடியது.

3.1.3.3 கணினிசார் பேரகராதிகள் (Computational lexicons)

மைய1980களில் கையால் அறிவு அடிப்படைகளை பெரிய அளவில் உருவாக்கும் பல முயற்சிகள் தொடங்கப்பட்டன [(எடுத்துக்காட்டாக wordnet/சொல்வலை (Miller et al 1990; Fellbaum 1998) Cyc (Lenat and Guha, 1990), ACQUILEX (Briscoe 1991), COMLEX (Grishman et al, 1994, Macleod et al)]. பொருண்மை சொற்களஞ்சியங்களை உருவாக்குவதற்கு இரண்டு அடிப்படை அணுகுமுறைகள் இருந்தன: அர்த்தங்கள் வெளிப்படையாகத் தரப்படும் பட்டியலிடும்/எண்ணிக்கையிடும் அணுகுமுறை மற்றும் ஆக்கமுறை அணுகுமுறை; இதில் தரப்பட்ட சொல்லுடன் தொடர்புள்ள பொருண்மைத் தகவல்கள் குறைவாகச்

சிறப்பிக்கப்பட்டுள்ளது மற்றும் துல்லியமான அர்த்தத் தகவல்களை ஆக்குவதற்கு ஆக்கமுறை விதிகள் பயன்படுத்தப்படுகின்றது.

பட்டியலிடும் பேரகராதிகள் (Enumerative lexicons)

பட்டியலிடும் அகராதிகளில் சொல்வலை (WordNet (Miller et al 1990; Fellbaum, 1998) தற்போது அதிகமாக அறியப்படுவதும் ஆங்கிலத்தில் சொற்பொருண்மை மயக்கநீக்கத்திற்கு அதிகமாகப் பயன்படுத்தப்படுவதும் ஆகும். பல மேற்கு மற்றும் கிழக்கு ஐரோப்பிய மொழிகளுக்குச் சொல்வலையின் பதிப்புகள் தற்போது உருவாக்கப்பட்டு வருகின்றன (Vossen, 1998; Sutcliffe et al 1996a and b).

சொல்வலை சொற்பொருண்மை மயக்கநீக்கத்திற்குச் சிறந்த மூலவளம் அல்ல. அடிக்கடி கூறப்படும் சிக்கல் சொல்வலையின் அர்த்த வேறுபடுகளின் அளவுக்கு அதிகமான கூறாக்கம் ஆகும்; அவைகள் பல மொழிப் பகுப்பாய்வுப் பயன்பாடுகளின் தேவைக்கும் அப்பாற்பட்டதாகும்.

ஆக்கமுறை அகராதிகள் (Generative lexicon)

பெரும்பாலான சொற்பொருண்மை மயக்கநீக்க ஆய்வுகள் அகராதியில் காணப்படும் பட்டியலிடப்பட்ட அர்த்த வேறுபாடுகளைச் சார்ந்து இருக்கின்றன. இருப்பினும் சொற்பொருண்மை மயக்கநீக்கத்திற்கு ஆக்கமுறை அகராதிகளைப் (Pustejovsky 1995) பயன்படுத்தும் தற்போதைய ஆய்வுகளும் உள்ளன; இதில் தொடர்புள்ள அர்த்தங்கள் (ஒப்புருமொழியத்திற்கு எதிரான ஒழுங்கான பல்பொருள் ஒருமொழியம்) பட்டியலிடப்படுவதில்லை; மாறாக உருவகம், ஆகுபெயர் போன்ற அர்த்த ஆக்கத்தில் உள்ள ஒழுங்கைப் பிணைக்கும் விதிகளிலிருந்து ஆக்கப்படுகின்றது. பியூடெலாரில் (Buitelaar 1997) சுருக்கமாகக் கூறப்படுவது போல் ஆக்கமுறைச் சூழலில் அர்த்த மயக்கநீக்கம் பொருண்மை அடையாளப்படுத்தத்தில் தொடங்குகின்றது; இது ஒரு சொல்லின் எல்லாச் சீராகத் தொடர்புபடுத்தப்பட்ட அர்த்தங்களைப் பிரதிபலிக்கும் கலவைத்தன்மையான அறிவு உருப்படுத்ததைச் சுட்டிக்காட்டுகின்றது; அதற்குப் பின்னர் நேர்வுகளைப் பற்றிய துல்லியமான அர்த்தத் தகவலைக் கொண்டிருக்கும் கருத்தாடல் சார்ந்த பொருள்கோளை ஆக்கக்கூடும். பியூடெலார் (Buitelaar 1997) குறை சிறப்பீடு செய்யப்பட்ட பொருண்மை அடையாளப்படுத்தத்திற்கு CORELEX-இன் பயன்பாட்டை விளக்குகின்றார் (பார்க்க Pusjejovsky et al 1995).

3.1.4 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைகள் (Corpus based methods)

3.1.4.1 அனுபவ நெறிமுறைகளின் வளர்ச்சி, வீழ்ச்சி, மீட்சி

பத்தொன்பதாம் நூற்றாண்டின் முடிவிலிருந்து தரவுத்தொகுதியின் மனிதப் பகுப்பாய்வு சொற்களின் மற்றும் எழுத்துக்களின் ஆய்வையும் மொழிப் பேறின் மற்றும் மொழி கற்பித்தலின் ஆய்வுக்குச் சொற்களின் மற்றும் சேர்ந்துவரும் சொற்களின் பட்டியலின் பிரித்தெடுப்புக்கும் சாத்தியமாக்கியது. இருபதாம் நூற்றாண்டின் முதல் அரைப் பகுதியிலிருந்து தரவுதொகுதிகள் மொழியியலில் பயன்படுத்தப்பட்டது.

தரவுத்தொகுதி மாதிரிகளின் வங்கியைத் தந்தது; இது எண்சார்ந்த மொழி மாதிரிகளின் உருவாக்கத்தைச் சாத்தியமாக்கியது; இவ்வாறு தரவுத்தொகுதிகளின் பயன்பாடு அனுபவ நெறிமுறைகளுடன் இணைந்து சென்றது.

தரவுத்தொகுதிகள் மற்றும் அனுபவ நெறிமுறைகள் இவற்றின் பயன்பாட்டிலிருந்து விலகிச் செல்லும் சூழலில் வெய்ஸ் (Weiss 1973) மற்றும் கெல்லெ மற்றும் ஸ்டோன் (Kelley and Stone 1975) கையால் அர்த்தம் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியிலிருந்து பொருண்மை மயக்கநீக்க விதிகளைக் கற்க இயலும் என்று நிரூபித்துக் காட்டினார்.

1980களில் தரவுத்தொகுதி மொழியியலின் ஆர்வம் மீளப்பெற்றது. தொழில் நுட்பத்தின் முன்னேற்றங்கள் முன்னர் சாத்தியமான தரவுத்தொகுதியைக் காட்டிலும் பெரிய தரவுத்தொகுதியின் உருவாக்கத்தையும் சேகரிப்பையும் சாத்தியமாக்கியது; பெரும்பாலும் புள்ளியியல் நெறிமுறைகளைப் பயன்படுத்தி இது புதிய மாதிரிகளின் உருவாக்கத்தைச் சாத்தியமாக்கியது. இந்த நெறிமுறைகள் பேச்சு ஆய்வில் முதலில் மீள்கண்டுபிடிப்பு செய்யப்பட்டது.

சொற்பொருண்மை மயக்கநீக்கத்தின் களத்தில் பிளாக் (Black 1988) 22 மில்லியன் டோக்கன்கள்/சொற்கள் கொண்ட தரவுத்தொகுதியைப் பயன்படுத்தி 2000 சொல்லடைவு வரிகளைக் கையால் அர்த்தத்திற்கு அடையாளப்படுத்தத்திய பின்னர் தீர்மானக் கிளைகள் (decision trees) அடிப்படையில் ஒரு மாதிரியை உருவாக்கினார். அதன் பிறகு பொருண்மை அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளிலிருந்து கண்காணிக்கப்பட்ட கற்றல் பல ஆய்வாளர்களால் பயன்படுத்தப்பட்டது. இருப்பினும் பெரும் தரவுத்தொகுதிகள் கிடைக்கும் நிலைமைக்குப் பின்னரும் இரண்டு முக்கியமான தடைகள் தரவுத்தொகுதிகளிலிருந்து சொல்சார் அறிவின் பேறை தடைசெய்தது: ஒரு பயிற்சி தரவுத்தொகுதியைக் கையால் அர்த்தத்தை அடையாளப்படுத்தும் சிரமங்கள் மற்றும் தரவு குறைவு.

3.1.4.2 தானியக்க அர்த்தம் அடையாளப்படுத்தல் (Automatic sense-tagging)

தரவுத்தொகுதியைக் கையால் அர்த்தத்திற்கு அடையாளப்படுத்துவது மிக விலையுயர்ந்ததாகும்; தற்போது அர்த்ததிற்கு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் மிக அரிதாகவே கிடைக்கின்றன. அர்த்தத்திற்கு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளின் உருவாக்கத்திற்குப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன: அண்மைக் காலத்தில் மொழியியியல் தரவுக் கூட்டிணைப்பு (Linguistic Data Consortium) பிரவுண் தரவுத்தொகுதியிலிருந்து கிட்டத்தட்ட 200,000 வாக்கியங்களின் தரவுத்தொகுதிகளை வினியோகித்தது. வால் ஸ்ட்ரீட் ஜர்னலில் (Wall Street Journal) 191 சொற்களின் எல்லா நேர்வுகளும் சொல்வலையின் அர்த்தங்களுடன் கையால் அடையாளப்படுத்தப்பட்டது. மேலும் பிரின்ஸ்டனில் உள்ள புலனறிவு அறிவு அறிவியல் பரிசோதனைக்கூடம்/காக்னிட்டிவ் சயன்ஸ் லபாரட்டரி (Cognitive Science Laboratory) பிரவுண் தரவுத்தொகுதியிலிருந்து 1000 சொற்களை அவற்றின் சொல்வலை அர்த்தங்களால் கையால் அடையாளப்படுத்துவதை நடைமுறைபடுத்தியது. இருப்பினும் புள்ளியியல் நெறிமுறைகளில் எடுத்துக்காட்டாகப் பயன்படுத்தப்படும் தரவுத்தொகுதியிலிருந்து மிகச் சிறிதாகும்.

ஸ்சூட்ச் (Schütze 1992) பயிற்சித் தரவுத்தொகுதியில் ஒவ்வொரு நேர்வையும் அடையாளப்படுத்துவதைத் தவிர்க்கும் ஒரு நெறிமுறையை முன்மொழிந்தார். 1001 எழுத்துச் சாளரத்திற்குள் எழுத்து நான்குகிராமைப் (letter fourgrams) பயன்படுத்தி அவரது நெறிமுறை முதலில் பனுவலில் உள்ள சொற்களைத் தானியக்கமாகத் திரட்டியது; ஒவ்வொரு இலக்குச் சொல்லும் ஒரு வெக்டாரால் உருப்படுத்தம் செய்யப்பட்டது; ஒரு அர்த்தம் ஒவ்வொரு நேர்வுக்கும் ஒதுக்கப்படுவதற்குப் பதிலாக ஒவ்வொரு கொத்துக்கும்/திரட்டுக்கும் ஒதுக்கப்பட்டது. இந்த அணுகுமுறை மனிதக் குறுக்கீட்டைக் குறைத்தது; இருப்பினும் ஒவ்வொரு பொருண்மை மயக்கம் உள்ள சொற்களுக்கும் நூறு அல்லது அதற்கு மேற்பட்ட நேர்வுகளின் பரிசோதனையை வேண்டியது. கொத்துக்களிலிருந்து ஆக்கப்பட்ட அர்த்தங்கள் எதனுடன் பொருந்துகிறது என்பதில் தெளிவில்லை; அவை தரவுத்தொகுதியிலிருந்து ஆக்கப்பட்டுள்ளதால் அவற்றைப் பிற ஒழுங்குமுறைகளில் பயன்படுத்த இயலாது.

பிரவுண் மற்றும் பிறர் (Brown et al 1991) மற்றும் கேல் மற்றும் பிறர் (Gale et al 1992a, 1993) பயிற்சித் தரவின் கையால் அடையாளப்படுத்தலைத் தவிர்க்க இருமொழியத் தரவுத்தொகுதியைப் பயன்படுத்துவதை முன்மொழிந்தனர். அவர்களின் முற்கூற்று (premise) தரப்பட்ட சொல்லின் வேறுபட்ட அர்த்தங்கள் மற்றொரு மொழியில் பெரும்பாலும்

வேறுபாட்டுடன் மொழிபெயர்ப்பு செய்யப்படுகின்றது (எடுத்துக்காட்டாக ஆங்கிலத்தில் *pen* என்பது பிரஞ்சு மொழியில் எழுதும் கருவி அர்த்தத்தில் *stylo* என்றும் அதன் வேலி அர்த்தத்தில் enclose என்றும் வரும்). இணையான வரிசைப்படுத்தப்பட்ட தரவுத்தொகுதியைப் (parallel aligned corpus) பயன்படுத்தி ஒரு சொல்லின் ஒவ்வொரு நேர்வின் மொழிபெயர்ப்பையும் அதன் அர்த்தத்தைத் தீர்மானிக்கத் தானியக்கமாகப் பயன்படுத்த இயலும். பல பொருண்மை மயக்கங்கள் இலக்கு மொழியில் தக்கவைக்கப்படுவதால் இந்த நெறிமுறைக்கு வரம்பு உண்டு (எடுத்துக்காட்டு பிரஞ்சு souris –ஆங்கில mouse).

3.1.4.3 தரவு அரிதாக இருப்பதை நேரிடுவது (Overcoming Data Sparseness)

தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகளில்/செயல்பாடுகளில் பொதுவாகக் காணப்படும் தரவு அரிதுச் (data scarce) சிக்கல் குறிப்பாகச் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகளில் கடினமானதாகும். முதலாவதாகப் பொருண்மைகளுக்குள் நிகழ்வெண்களில் அதிக அளவு ஏற்றதழ்வு இருப்பதன் காரணமாக ஒரு பல்பொருண்மைத் தன்மையான சொல்லின் எல்லாப் பொருண்மைகளும் உருப்படுத்தம் செய்யப்பட்டுள்ளது என்பதை உறுதி செய்ய வேண்டி அதிக அளவிலான பனுவல்கள் தேவைப்படும். எடுத்துக்காட்டாக, பிரவுண் தரவுத்தொகுதியில் (ஒரு மில்லியன் சொற்கள்) ஒத்தறி அடிப்படையில் சாதாரணமான சொல்லான *ash* என்பது எட்டு தடவை மட்டுமே நேர்ந்துள்ளது; மற்றும் அதன் *tree* என்ற பொருண்மையில் ஒரு தடவை மட்டுமே நேர்ந்துள்ளது. *ashes* = *remains of cremated body* 'எரிக்கப்பட்ட உடலின் எச்சம்' என்ற பொருண்மை LDOCE (Longman's Dictionary of Contemporary English) மற்றும் OALD (Oxford Advanced Learner's Dictionary) போன்ற கற்பவர்களின் அகராதில் உட்படுத்தப்படும் அளவுக்குச் சாதாரணமானதாக/பொதுவானதாக இருந்தாலும் அவற்றில் வரவில்லை; மற்றும் CED (Chamber's English Dictionary) போன்ற பல அன்றாட அகராதிகளில் ஒரு டசன் அல்லது அதற்கு மேல் பொருண்மைகள் காண்பது கிட்டத்தட்ட சாத்தியமில்லாததாகும். மேலும் தரப்பட்ட ஒரு பல்பொருண்மைத் தன்மையான சொல்லுக்குப் பல சாத்தியமான சேர்ந்துவரும் நேர்வுகளை மிகப் பெரிய தரவுத் தொகுதியில் கூட காண்பது அரிதாகும், அல்லது அவற்றைச் சிறப்பானதாகக் கருதவியல்லாத அளவுக்கு அரிதாக வரும்.

எளிதாக்குதல்/நிரப்பாக்குதல் (smoothing) மிகக்குறைவாக நேரும் நிகழ்வுகளின் சிக்கலை நேரிடப் பயன்படுத்தப்படுகின்றது; குறிப்பாக உற்று நோக்கப்படாத நிகழ்வுகள் பூஜியச் சாத்தியக் கூறு உள்ளதாகக் கருதப்படவில்லை என்பதை உறுதி செய்ய வேண்டி பயன்படுத்தப்படுகின்றது. மிக

அறியப்பட்ட நிரப்பாக்கும் நெறிமுறைகள் டர்னிங்-குட் (Good 1953) என்பதாகும்; இது நிகழ்வுகளின் இருபெயர்சார் வினியோகத்தைக் (binominal distribution) கருதுகோள் செய்கின்றது; மற்றும் ஜெல்னெக் மற்றும் மெர்சர் (Jelinek and Mercer 1985) என்போரின் நெறிமுறைப் பயிற்சித் தரவுத்தொகுதின் வேறுபடுத்தப்பட்ட துணைப் பாகங்களின் மேல் மதிப்பீடு செய்யப்பட்ட காரணிகளை இணைக்கின்றது. இருப்பினும் இந்த நெறிமுறைகள் *ash-cigarette* மற்றும் *ash-room* போன்ற ஒரே நிகழ்வெண்ணைக் கொண்ட நிகழ்வுகளுக்கு இடையில் உள்ள வேறுபாட்டை நிறுவுவதைச் சாத்தியமாக்குவதில்லை. சர்ச் மற்றும் கேல் (Church and Gale 1991) என்போர் பைகிராம்களின் (இலக்குச் சொல்லுக்கு அடுத்து வரும் இருசொற்களின் ஒரு வகைக் கணிப்பு) மதிப்பீட்டிற்கு நெறிமுறைகளை மேம்படுத்தும் ஒரு வழியை முன்மொழிந்துள்ளனர்; இதைச் சேர்ந்துவருகை நேர்வுகளுக்கும் நீட்சி செய்யலாம்: அவர்கள் பைகிராமை உருவாக்கும் தனிச் சொற்களின் நிகழ்வெண்ணைக் கருத்தில் கொள்கின்றனர்; மற்றும் ஒவ்வொரு சொல்லும் பிற சொற்களிலிருந்து சுதந்திரமாக வருகின்றன என்று பொருள்கோள் செய்கின்றனர். இருப்பினும் இந்தப் பொருள்கோள் சேர்ந்துவருகை நேர்வு அடிப்படையிலான பொருள்மயக்க நீக்கத்தின் பொருள்கோளிடமிருந்து முரண்படுகின்றது; இது சில சேர்க்கைகள் பிற சேர்க்கைகளைவிடக் கூடுதல் சாத்தியமானது என்று சரியாகக் கருதுகின்றது.

வகுப்பு அடிப்படையிலான மாதிரிகள் (Class-based models) ஒரு பொது வகைப்பாட்டைச் சார்வதாகக் கருதப்படும் சொற்களின் வகுப்புகளின் உற்று நோக்குகளை இணைத்துப் பெறப்படும் மிகச் சிறந்த மதிப்பீடுகளைப் பெறுவதற்கு முயலுகின்றது. பிரவுண் மற்றும் பிறர் (Brown et al. 1992), பெரைரா மற்றும் டிஷ்பை (Pereira and Tishby 1992), டிஷ்பை மற்றும் லீ (Pereira, Tishby and Lee 1993) என்போர் தரவுத்தொகுதியின் வினியோகத் தனித்தன்மைகளிலிருந்து வகுப்புகளை ஆக்கும் நெறிமுறைகளை முன்மொழிகின்றனர்; பிற படைப்பாளிகள் வகுப்புகளை வரையறை விளக்கம் செய்வதற்காகப் புறத் தகவல் மூலங்களைப் பயன்படுத்துகின்றனர்: ரெஸ்னிக் (Resnik 1992) சொல்வலையின் வகைப்பாட்டியலைப் பயன்படுத்துகின்றார்; யரொவ்ஸ்கி (Yarowsky 1992) ரோஜட் தெசாரசின் (Roget's Thesaurus) வகைபாடுகளைப் பயன்படுத்துகின்றார்; ஸ்லாடர் (Slator 1992) மற்றும் லிட்டி மற்றும் பைக் (Liddy and Paik 1993) LDOCEஇல் எழுவாய் குறியங்களைப் பயன்படுத்துகின்றார்; லக் (Luk 1995) LDOCE-இன் வரையறை விளக்கங்களிலிருந்து உருவாக்கப்பட்ட கருத்துருக் குழுமங்களைப் பயன்படுத்துகின்றார். வகுப்பு-அடிப்படையிலான நெறிமுறைகள் தரவு அரிதின்

சிக்கல்களுக்குப் பகுதி விடை அளிக்கின்றது மற்றும் முன்-அடையாளப்படுத்தப்பட்ட தரவின் தேவையை நீக்குகின்றது. இருப்பினும் ஒரே வகுப்பில் வரும் எல்லாச் சொற்களும் ஒரே மாதிரி நடத்தையைக் கொண்டிருக்கும் என்ற கருதுகோள் மிக வலுவானதாகும். இருப்பினும் இந்த நெறிமுறைகளால் தகவல் இழப்பு உள்ளது. எடுத்துக்காட்டாக *residue* என்பது சொல்வலையில் *ash* என்பதன் உள்ளடக்கு மொழியாகும் (hypernym); அதன் உள்ளடங்கு மொழி (hyponym) {ash, cotton (seed) cake, dottle} என்ற வகுப்பை உருவாக்கும். வெளிப்படையாக, இச்சொற்களின் குழுமத்தின் உறுப்பினர்கள் சூழலில் வேறுபாடாக நடத்தைசெய்கின்றன, ஆனால் இக்குழுமத்தில் பிற சொற்களுடன் சிறிது உறவுடனோ அல்லது அல்லது உறவு ஏதுமின்றியோ இருக்கின்றன.

ஒற்றுமை அடிப்படையிலான நெறிமுறைகள் (Similarity-based methods): டாகன், மார்கஸ் மற்றும் மர்கொவிட்ச் (Dagan, Marcus and Markovitch 1993), டாகன், பெரைரா மற்றும் லீ (Dagan, Pereira and Lee 1994), கிரிஷ்மான் மற்றும் ஸ்டெர்லிங் (Grishman and Sterling 1993) என்போர் சொற்களைக் குறிப்பிட்ட வகுப்புகளாக மீள்குழுமம் செய்யாமல், ஒற்றுமையுள்ள சொற்களின் உற்றுநோக்குகளைக் குழுமம் செய்யும் கருத்தைப் பயன்படுத்தினர். ஒவ்வொரு சொல்லும் ஆற்றலுள்ள வெவ்வேறுபட்ட ஒற்றுமையுள்ள சொற்களின் குழுமத்தைக் கொண்டிருக்கும். பிரவுண் மற்றும் பிறரின் (Brown et al. 1992) பல வகுப்பு அடிப்படையிலான நெறிமுறைகளைப்போல் ஒற்றுமை அடிப்படையிலான நெறிமுறைகள் சேர்ந்துவருகையின் அமைப்பொழுங்குகளுக்கு (patterns) இடையில் உள்ள ஒரு ஒற்றுமை அளவுகோலைப் பயன்படுத்துகின்றது. டாகன், மர்கஸ் மற்றும் மார்கோவிட்ச் (1993) பின்வரும் எடுத்துக்காட்டைத் தருகின்றனர்: (chapter, describes) என்ற இணை அவர்களின் தரவுத்தொகுதியில் தோன்றவில்லை; இருப்பினும், chapter என்பது describes என்பதுடன் இணையாக வரும் book, introduction, section என்பனவற்றுடன் ஒற்றுமையுள்ளது. மாறாக books, documentation, manuals என்பன book என்ற சொல்லுடன் ஒற்றுமை உடையன. டாகன், மார்காஸ், மார்கொவிட்ச் (Dagan, Marcus mand Markovitch 1993) என்போரின் மதிப்பீடு, ஒற்றுமை அடிப்படையிலான நெறிமுறையானது வகுப்பு அடிப்படையிலான நெறிமுறைகளைவிட நன்றாகச் செயல்புரியும் எனக்காட்டுகின்றது. கரொவ் மற்றும் எடெல்மன் (Karov and Edelman) என்போர் கற்கும் நிலையில் திரும்பச்செய்யும் செயற்பாங்கின் மூலம் ஒற்றுமை அடிப்படையிலான நெறிமுறையின் நீட்சியை முன்மொழிகின்றனர்; இது நான்கு பரிசோதனை சொற்களில் 92% துல்லியத்தைத் தருகின்றது; இது இதுவரை இது குறித்த இலக்கியத்தில் மிகச் சிறந்த முடிவு என்று

கூறப்படுவதுடன் ஓரளவுக்கு ஒற்றுமையைக் காட்டுகின்றது. நூற்றுக்கணக்கான எடுத்துக்காட்டுகளை வேண்டும் பிற நெறிமுறைகளைப் போலல்லாமல் இம்முறையில் பயிற்சித் தரவுத்தொகுதி ஒவ்வொரு சொல்லுக்கும் சிறிதளவு எடுத்துக்காட்டுகளையே கொண்டிருக்கின்றது என்று வரும்போது இந்த முடிவுகள் குறிப்பாக மனதில் பதியத்தக்கது.

3.2 திறந்த சிக்கல்கள் (open problems)

நாம் குறிப்பிட்ட நெறிமுறைகளுடன் தொடர்புள்ள தற்போதைய சொற்பொருள் மயக்கநீக்க ஆய்வில் எதிர்கொண்ட பல்வேறு சிக்கல்களைப் பற்றி முன்னர் பார்த்தோம். இப்பகுதியில் எல்லாச் சொற்பொருள் மயக்கநீக்க அணுகுமுறைகளும் எதிர்கொள்ள வேண்டிய விவாத விசயங்களும் சிக்கல்களும் விளக்கப்படுகின்றது.

3.2.1 சூழலின் பங்களிப்பு (the role of context)

சூழல் ஒரு பல்பொருண்மைத்தன்மையான சொல்லின் (polysemous word) பொருண்மையை அடையாளம் காணும் ஒரே வழியாகும். எனவே, பொருண்மை மயக்கநீக்கத்தின் எல்லாச் செயல்பாடுகளும் இலக்குச் சொல்லின் பொருண்மை மயக்கநீக்கத்திற்குப் பயன்படுத்தப்படுவதற்குத் தகவல்களைத் தரவேண்டி அச்சொல்லின் சூழலைச் சார்ந்திருக்கின்றது. தரவு-இயக்க நெறிமுறைகளுக்குச் (data-driven methods) சூழலானது பொருண்மை மயக்கநீக்கத்தை நிறைவேற்ற வேண்டி தற்போதைய சூழலுடன் ஒப்பிடப்படும் முந்தைய அறிவையும் தருகின்றது.

பொதுவாகக் கூறினால் சூழல் இரு வழிகளில் பயன்படுத்தப்படுகின்றது:

- சொற்களின் பை அணுகுமுறை (bag-of-words approach): சூழல் என்பது எதாவது ஒரு சாளரத்தில் இலக்குச் சொல்லைச் சுற்றி இருக்கும் மற்றும் இலக்குச் சொல்லுடன் அவற்றின் உறவுகள் தூரம், இலக்கண உறவு போன்றவற்றின் அடிப்படையில் கருத்தில் கொள்ளப்படாமல் ஒரு குழுவாகக் கருதப்படுகின்ற சொற்கள் ஆகும்.
- உறவுசார் தகவல்: சூழல் என்பது இலக்குச் சொல்லிலிருந்து உள்ள தூரம், தொடரியல் சார் உறவுகள் (syntactic relations), விருப்பத் தேர்வுகள் (sectional preferences), எழுத்துசார் தனித்தன்மைகள் (orthographic properties), தொடர்சார் சேர்ந்துவருகை (phrasal collocation), பொருண்மை வகைபாடுகள் போன்றவற்றை உள்ளடக்கிய சில உறவுகள் அடிப்படையில் கருதப்படும்.

பொருண்மைத் தேர்வுக்கு நுண்சூழல் (microcontext), தலைப்புச் சூழல் (topical context), பொருட்புலம் (domain) என்பன பங்களிப்பு செய்யும்; ஆனால் ஒத்தறி பங்களிப்புகளும் (relative

roles) வேறுபட்ட சூழலிருந்து கிடைக்கும் தகவல்களின் முக்கியத்துவமும் அவற்றிற்கிடையிலான உறவுகளும் (interrelations) நன்றாக அறியப்படவில்லை. மிகக் குறைவான ஆய்வுகளே இம்மூன்று வகைத் தகவல்களைப் பயன்படுத்துகின்றன; மற்றும் தற்போதைய ஆய்வுகள் நுண்சூழலில் கூடுதல் கவனக்குவிப்பு செய்கின்றன.

3.2.1.1. நுண்சூழல் (Microcontext)

பெரும்பாலான பொருண்மை மயக்கநீக்கச் செயல்பாடுகள் சொற்பொருண்மை மயக்கநீக்கத்திற்கு முதன்மைத் தகவல் மூலமாக ஒரு சொல் நேர்வின் வட்டாரச் சூழலைப் பயன்படுத்துகின்றது. வட்டார அல்லது "நுண்" சூழல் பொதுவாக ஒரு பனுவலில் அல்லது கருத்தாடலில் ஒரு சொல் நேர்வைச் சுற்றியிருக்கும் சொற்களின் ஏதாவது ஒரு சாளரமாகக் கருதப்படுகின்றது; இது இலக்குச் சொல் தோன்றும் சூழலின் சில சொற்களிலிருந்து மொத்த வாக்கியத்தையும் உள்ளடக்கும்.

சூழல் என்பது தூரம், தொடரமைப்பு அல்லது பிற உறவுகளைக் கருதில் கொள்ளாமல் இலக்கின் ஏதாவது சாளரத்தில் வரும் எல்லாச் சொற்களாகவும் அல்லது எழுத்துகளாகவும் கருதப்படும். வெயிசின் (Weiss 1973) ஆய்வு போன்ற தொடக்ககாலத் தரவுத்தொகுதி அடிப்படையிலான செயல்பாடு இவ்வணுகுமுறையைப் பயன்படுத்தியது. செயலூக்கத்தைப் பரப்பும் (spreading activation) மற்றும் அகராதி அடிப்படையிலான (dictionary based) அணுகுமுறைகளும் ஒரு சாளரத்தில் நேர்வதன் அடிப்படை தவிர பிற எந்த அடிப்படையிலும் சூழல் உள்ளீட்டைப் பொதுவாக வேறுபடுத்துவதில்லை. ஸ்சுட்செயின் (Schutze's) திசையன்/வெக்டர் இட நெறிமுறை (vector space method) அடுத்துவரும் தகவலை (adjacency) புறக்கணிக்கும் அணுகுமுறைக்கு ஒரு அண்மைக்கால எடுத்துக்காட்டாகும். ஒட்டுமொத்தமாகச் செற்களின் பை அணுகுமுறை வினைகளைவிட பெயர்களுக்கு நன்றாகச் செயல்படுவதாகத் தெரிகின்றது; இவணுகுமுறை பொதுவாக உறவுகளைக் கருத்தில்கொள்ளும் பிற அணுகுமுறைகளைக் காட்டிலும் குறைந்த செயல் திறம் உள்ளதாகவும் இருக்கின்றது. இருப்பினும் யாரொவ்ஸ்கியின் (Yarowsky 1992) ஆய்வில் நிரூபிக்கப்பட்டுள்ளதுபோல் இவணுகுமுறை கூடுதல் சிக்கலான (கலவைத்தன்மையான) செயற்பாங்கை வேண்டும் அணுகுமுறைகளைக் காட்டிலும் சிக்கனமானது மற்றும் சில பயன்பாடுகளில் போதுமான அளவு பொருள்மயக்கநீக்கத்தை நிறைவேற்ற இயலும். கீழே நாம் பிற காரணிகளைப் பரிசோதிக்கப் போகின்றோம்.

தூரம் (distance): வீவரின் நினைவுக்குறிப்பிலிருந்து (Weaver's memorandum) தொடக்ககாலத்திலிருந்தே பொருண்மை மயக்கநீக்கத்திற்கு இலக்கைச் சுற்றி வரும் சில சொற்களின் சூழலைப் பரிசோதிக்கும் கருதுச்சாயல் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாட்டிற்கு அடிப்படையாகும் என்பது வெளிப்படையாகும்: இது இயந்திர மொழிபெயர்ப்புச் செயல்பாடு, பொருளடக்கப் பகுப்பாய்வு (content analysis) , செயற்கை அறிவு பொருண்மை மயக்கநீக்கம் (AI-based disambiguation), அகராதி அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் (dictionary based WSD), மேலும் அண்மைக்கால புள்ளியியல்சார், நரம்பு வலையமைப்பு (nueural network) மற்றும் குறியீடுசார் இயந்திரக் கற்றல் (symbolic machine learning) அணுகுமுறைகள் என்பனவற்றிற்கு அடிப்படையாகும். இருப்பினும் கப்லானின் (Kaplan 1950) தொடக்ககாலப் பரிசோதனைகளைப் பின்பற்றி N என்பதன் அனுகூலமான மதிப்பைக் கருதில் கொள்ளும் வீவரின் கேள்விக்குப் பதில் அளிக்கவேண்டி சில ஒழுங்கான முயற்சிகள் மேற்கொள்ளப்பட்டன. இதில் குறிப்பிடத்தக்க விதிவிலக்கு, சொற்பொருள் மயக்கநீக்கத்திற்கு மிக நம்பிக்கையான இரண்டு சூழல்கள் முக்கியம் என்ற கப்ளானின் கண்டுபிடிப்பைச் சரிபார்த்த சௌயெகா மற்றும் லுசிக்னன் (Choueka and Lusignan 1985) என்போர்களின் ஆய்வு ஆகும்; 1-சூழல்கள் 10 நேர்வுகளில் 8 என்ற நிலையில் நம்பத்தகுந்ததாகும். இருப்பினும் இக்கண்டுப்பிடிப்புகளுக்கு மாறாக N என்பதன் மதிப்பு சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகளின் போக்கில் இடுகுறித்தன்மையாகத் தொடர்ந்து வேறுபடலாயிற்று.

யரோஸ்கி (Yarosky 1993, 1994a, 1994b) 1-சூழல்கள், k-சூழல்கள் மற்றும் பக்கக் கிளைகள் (offsets) -1 மற்றும் -2, -1 மற்றும் +1, மற்றும் +1 மற்றும் +2 இவற்றை உள்ளடக்கிய நுண்சூழலின் வேறுபட்ட சாளரங்களைப் பரிசோதித்தார்; மேலும் பொருண்மை மயக்கநீக்கத்திற்கு மிக நம்பத்தகுந்த சான்றைக் கண்டுபிடிக்க ஒரு log-சாத்தியக்கூறு விகிதத்தைப் பயன்படுத்தி வகைப்படுத்தினார். யரோவ்ஸ்கி k என்பதன் அனுகூலமான மதிப்பு பொருண்மை மயக்கத்தின் வகையைப் பொறுத்து மாறுபடுகின்றது என்ற உற்றுநோக்கைச் செய்தார்: அவர் வட்டார மயக்கங்கள் (local ambiguities) k = 3 அல்லது 4 என்பதன் ஒரு சாளரத்தை மட்டும் வேண்டும், அதே சமயம் பொருண்மை அல்லது தலைப்பு அடிப்படையிலான மயக்கங்கள் 20-50 சொற்களைக் கொண்ட சாளரத்தை வேண்டும் என்று கூறினார். ஒரு தனிப்பட்ட ஏற்றமிக்க அளவு எடுத்துக் கூறப்படவில்லை; இது வேறுபட்ட மயக்கம் உள்ள சொற்களுக்கு வேறுபட்ட தூர உறவுகள் கூடுதல் திறன் வாய்ந்ததாய் அமையும் என்று கருத்து

தெரிவிக்கின்றது. மேலும் யரோஸ்கி பிற தகவல்களையும் பயன்படுத்தியதன் காரணமாகச் சாளர வடிவ அளவின் விளைவை மட்டும் தனியாகப் பிரிப்பது கடினமாகும். லீகாக், சொடொரொ, மில்லர் (Leacock et al 1998) என்போர் ±3 திறந்த-வகுப்பு சொற்களின் (open-class words) வட்டார சாளரத்தைப் பயன்படுத்துகின்றனர்; அவர்கள் முந்தையப் பரிசோதனைகளில் இவ்வெண் ஏற்றமிகுந்த நிறைவேற்றதைக் காட்டியது என்று வாதிடுகின்றர்.

சேர்ந்துவருகை (collocation): "சேர்ந்துவருகை" என்ற சொல் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாட்டில் பலவிதமாகப் பயன்படுத்தப்படுகின்றது. இச்சொலை ஜெ.ஆர். ஃபிர்த் 1951-இல் "Modes of meaning" என்ற தமது கட்டுரையின் மூலம் பலரும் அறியச்செய்தார்: ass என்பதன் பொருண்மைகளில் ஒன்று அதற்கு உடனடியாக முன்வரும் "you silly…" என்பதுடன் அதன் வழக்கமான சேர்ந்துவருதல்" ஆகும். அவர் சேர்ந்துவருதல் என்பது உடன்வருதல் (co-occurrence) என்பதல்ல, "வழக்கமானது" அல்லது "சாதாராணமானது" என்பதாகும் என்று வலியுறுத்திக் கூறினார். சேர்ந்துவருகை என்பதற்கு ஹாலிடேயின் (Haliday 1961) வரையறைவிளக்கமான "இது சொற்களின் உறுப்பமைவுச் சேர்க்கையாகும்; இது பனுவல் அடிப்படையில்அளவிடத் தக்கது; இது x என்ற ஒன்றிலிருந்து n விலகலில் (n சொற்களின் தூரம்) a, b, c… என்பனவைகள் நேரும் சாத்தியமாகும்" என்பது கணினி அடிப்படையில் கூடுதல் செயல்புரியும்.

இவ்வரையறை விளக்கம் அடிப்படையில் ஒரு தனிச்சிறப்பான சேர்ந்துவருகை (significant collocation) என்பதைச் சொற்களுக்கிடையிலுள்ள உறுப்பமைவுத் தொடர்பாகும் (syntagmatic association) என்று வரையறை விளக்கம் செய்யலாம்; இதில் a, b, c… என்பனவற்றுடன் x சேர்ந்துவரும் சாத்தியம் தற்செயல் நேர்வைவிடக் கூடுதலாகும் (Berry-Rogghe 1973). இப்பொருண்மையில் தான் பெரும்பாலான சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகள் இக்கலைச் சொல்லைப் பயன்படுத்துகின்றன. சேர்ந்துவருகைகள் (collocations) பிற உடன்வருகைகளிலிருந்து (co-occurrences) வேறுபடுத்தப்பட்டு கையாளப்படுகின்றன என்பதற்குச் சில உளவியல் சான்றுகள் உள்ளன. எடுத்துக்காட்டாகக் கிந்த்ச் மற்றும் ம்ரோஸ் (Kintsch and Mross 1985) சோதனைச் சொற்களுடன் (i.e. iron-steel, அவர்கள் இதை உடன்வருகைச் சூழல் (associative context) என்கின்றனர்)) அடிக்கடி சேர்ந்துவருகைகள் செய்யும் முதன்மைச் சொற்கள் (priming words) சொல்சார் தீர்மானச் செயல்களில் (lexical decision tasks) இச்சோதனைச் சொற்களை ஊக்குவிக்கும் என்று காட்டினர். மறுதலையாக

மையக்கருத்துச் சூழலில் (thematic context) [(அதாவது plane-gate போன்று சூழ்நிலை (situation), காட்சி (scenario) அல்லது உரை (script) என்பனவற்றால் தீர்மானிக்கப்படும் உறவுகள்)] முதன்மைச் சொற்கள் ஆய்வுப்பொருள்களின் சொல்சார் தீர்மானங்களை எளிதாக்காது (பார்க்க Fischler 1977, Seidenberg et al 1982, De Groot 1983, Lupker 1984).

யரொவ்ஸ்கி (Yarowsky 1983) சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகளில் சேர்ந்துவருகைகளின் பயன்பாட்டை வெளிப்படையாகக் கூறுகின்றது; ஆனால் அவர் தமது குறிக்கோளுக்கு வேண்டி "சில வரையறை விளக்கம் செய்யப்பட்ட உறவுகளில் இரு சொற்களின் உடன்வருகை" என்ற வரையறை விளக்கத்தை ஏற்றுக்கொள்கின்றார். முன்னர் பார்த்தபடி அவர் பல்வேறுபட்ட தூர உறவுகளைப் பரிசோதிப்பதுடன் சொல்வகைப்பாட்டால் அண்மை வருகையையும் (adjacency) கருத்தில்கொள்கின்றார் (எ.கா. இடதுபக்கத்தில் முதல் பெயர்). அவர் இரு பொருண்மை மயக்க நேர்வில் ஒரு சேர்ந்துவருகைக்கு ஒரு பொருண்மை என அமைகிறது என்று நிர்ணயிக்கின்றார்; அதாவது ஒரு சேர்ந்துவருகையின் போது, ஒரு சொல் 90-99% சாத்தியத்துடன் ஒரே பொருண்மையில் பயன்படுத்தப்படுகின்றது.

தொடரியல் உறவுகள் (Syntactic Relations): இரல் (Earl 1973) தொடரியலை இயந்திர மொழிபெயர்ப்பின் பொருண்மை மயக்கநீக்கத்திற்கு மட்டுமே பயன்படுத்துகின்றார். இன்றுவரையிலான பெரும்பாலான சொற்பொருள் மயக்கநீக்கத்திற்குத் தொடரியல் தகவல் பிற தகவலுடன் இணைந்து பயன்படுத்தப்படுகின்றது. தேர்வுக்கட்டுப்பாடுகளின் (selectional restrictions) பயன்பாடானது முழுப் பகுத்துக்குறித்தல், பொருண்மையியல் சார் வலையமைப்புகள் (semantic networks), தேர்வு விருப்பங்கள் (selectional preferences) போன்றவற்றைச் சார்ந்திருக்கும் செயற்கை அறிவு அடிப்படையிலான (AI-based) செயல்பாட்டில் கூடுதலாகச் செல்வாக்குபெற்றுள்ளது (Hayes 1977a, 1977b; Wilks 1973 and 1975b; Hirst 1987). பிற செயல்பாடுகளில், தொடரியல் சேர்ந்துவருகைத் தகவல்களுடன் இணைத்துக் கொள்ளப்படுகின்றது: கெல்லி மற்றும் ஸ்டொன், டால்க்ரென், அட்கின்ஸ் (Kelley and Stone 1975, Dahlgren 1988, Atkins 1987) என்போர் சேர்ந்துவருகைத் தகவல்களை அடைகொளி அடைகள் (determiners), மாற்றுப்பெயர்கள், பெயர் நிரப்பிகள் (noun complements), முன்னுருபுகள் (prepositions), எழுவாய்-வினை (subject-verb) மற்றும் வினை-செயப்படுபொருள் (verb-object) உறவுகள் என்பனவற்றின் இருப்பை அல்லது இல்லாமையை நிர்ணயிக்கும் விதிகளுடன் இணைக்கின்றனர்.

சமீபத்தில் ஆய்வாளர்கள் ஆழமில்லா அல்லது பகுதிப் பகுப்பாய்வை (shallow or partial parsing) பயன்படுத்தி கலவைத்தன்மையான பகுப்பாய்வை விளக்குகின்றனர். பெயர்மீதான பொருண்மை மயக்கநீக்கச் செயல்பாட்டில், ஹியர்ஸ்ட் (Hearst 1991) பனுவல்களைப் பெயர்த்தொடர்கள், பின்னுருபுத் தொடர்கள் மற்றும் வினைக் குழுக்கள் என்பவைகளாகப் பிரித்து எடுத்துவிட்டுப் பிற எல்லாத் தகவல்களையும் தவிர்க்கின்றார். அவர் இலக்குச் சொல்லிருந்து ± 3 தொடர்ப் பகுதிகளுக்குள் (phrase segments) வரும் எல்லாச் சொற்களையும் ஆய்ந்துப் பின்னர் பெரிய எழுத்தாக்கம் (capatitalization) போன்ற பிற சான்றுகளுடன் தொடரியல் சான்றுகளை இணைக்கின்றார். யரொவ்ஸ்கி (Yarowsky 1993) தொடரியல் வகைப்பாடு அடிப்படையிலான பல்வேறு நடத்தைகளை நிர்ணயிக்கின்றார். எடுத்துக்காட்டாக, வினைகள் அவற்றின் எழுவாய்களைவிட செயப்படுபொருள்களிலிருந்து பொருண்மைமயக்கம் நீக்கும் தகவல்களைப் பெறுகின்றன; பெரும்பான்மையான எல்லா பொருண்மை மயக்கநீக்கத் தகவல்களையும் அவை அடைசெய்யும் பெயர்களிலிருந்து பெறுகின்றன; பெயர்கள் நேரடியாக அண்மையிலுள்ள பெயரடைகளாலோ பெயர்களாலோ பொருண்மை மயக்கநீக்கம் செய்யப்படுகின்றன. அண்மைக்காலப் பொருண்மை மயக்கநீக்கச் செயல்பாட்டில் தொடரியல் தகவல் பொரும்பான்மையும் பிற வகையிற்படும் தகவல்களுடன் இணைந்து பயன்படுத்தபடும் சொல்வகைப்பாடாகும் (McRoy 1992; Bruce and Wiebe 1994; Leacock, Chodorow, and Miller 1998). இருப்பினும் இன்றுவரை இலக்குச் சொல்லின் வேறுபட்ட வகைகளுக்கு வேண்டிய வேறுபட்ட தகவல் வகைகளுக்குப் பங்களிப்பு செய்யும் முறையான ஆய்வு மிகக் குறைவானதே. இதுதான் சொற்பொருண்மை மயக்கநீக்கச் செயற்பாட்டின் அடுத்த தேவையான நடவடிக்கையாக இருக்கும்.

3.2.1.2 தலைப்பு சார் சூழல் (topical context)

தலைப்புசார் சூழல் (topical context) பெரும்பாலும் பல வாக்கியங்கள் வரும் ஒரு சாளத்திற்குள் ஒரு சொல்லின் தரப்பட்ட பொருண்மையுடன் சேர்ந்துவரும் துணைச் சொற்களை (substantive words) உட்படுத்தும். 1950-களின் தொடக்கத்திலிருந்து பொருண்மை மயக்கச் செயல்பாடுகளில் பங்களிப்பு செய்த நுண்சூழலைப் போலல்லாமல், தலைப்புசார் சூழல் மிகக் குறைவான நிலைபாட்டிலேயே பயன்படுத்தப்பட்டுள்ளது. தலைப்புசார் சூழல் சார்ந்த நெறிமுறைகள் பனுவலில் உள்ள மிகைகளை (redundancy) பயன்படுத்தும்; அதாவது தரப்பட்ட ஒரு தலைப்பில் அமைந்த ஒரு பனுவல் முழுவதும் பொருண்மை அடிப்படையில் உறவுள்ள சொற்கள் திரும்பத் திருப்பப் பயன்படுத்தபடும். இப்படியாக *base* என்ற சொல் பொருண்மை

மயக்கம் உள்ளதாகும்; ஆனால் *pitcher*, *ball* என்ற சொற்களைக் கொண்டிருக்கும் ஒரு ஆவணத்தில் அதன் (*base* என்பதன்) தோற்றம் அச்சொல்லின் தரப்பட்ட பொருண்மையைப் பிரித்தெடுக்கும்; (மற்றும் பொருண்மை மயக்கம் உள்ள பிற சொற்களையும் பிரித்தெடுக்கும்). தலைப்புச் சூழலை உட்படுத்தும் செயல்பாடுகள் தனிச்சிறப்பாகச் சொற்களின் பை அணுகுமுறையைப் (bag-of-words approach) பயன்படுத்தும்; இதில் சூழலில் உள்ள சொற்கள் ஒழுங்குபடுத்தப்படாத குழுமமாகக் கருதப்படும்.

பல ஆண்டுகளாகத் தகவல் மீட்புக் களத்தில் தலைப்புச் சூழலின் பயன்பாடு விவாதிக்கப்படுகின்றது (Antthony 1954; Salton 1968). அண்மைக்காலத்தில் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகள் தலைப்புச் சூழலைப் பயன்படுத்துகின்றன: யரோவ்ஸ்கி (Yarowsky 1992) ரோஜஸ்டின் சொற்களஞ்சியத்தைப் (Rogest's thesarurus) பயன்படுத்தும் அவரது பரிசோதனையில் தொடர்புடைய சொற்களின் வகுப்புகளை உருவாக்கவும் மற்றும் பல்பொருண்மை இலக்கைச் சுற்றியுள்ள சூழலாக உருவாக்கவும் 100-சொல் சாளரத்தைப் பயன்படுத்துகின்றார். வூரீஸ், லீகாக் மற்றும் டோவல் (Voorhees, Leacock and Towell 1995) ஆகியோர் இரு வாக்கியச் சாளரத்தைப் பயன்படுத்தி பல புள்ளியியல் நெறிமுறைகளை பரிசோதனை செய்கின்றனர்; லீகாக், டோவல் மற்றும் வூரீஸ் (Leacock Towell and Voorhees 1993, 1996) ஆகியோர் இதுபோன்று சொற்பொருண்மை மயக்கநீக்கத்திற்குத் தலைப்புச் சூழலை ஆய்ந்தனர். கேல், சர்ச் மற்றும் யரோஸ்கி (Gale, Church and Yarowsky 1993) ஆகியோர் ±50 சொற்களின் சூழலைப் பார்த்து இலக்குச் சொல்லுக்கு மிக அருகில் வரும் சொற்கள் பொருண்மை மயக்கநீக்கத்திற்குப் பெரும்பான்மையும் பங்களிப்பு செய்வதைக் குறிப்பிட்டனர்; அவர்கள் இலக்கைச் சுற்றி வரும் ±6 (நுண்சூழலைக் கருத்தில் கொள்ளும் போது ஒரு தனிச்சிறப்பான அளவு) என்பதிலிருந்து ±50 சொற்களுக்கு சூழலை விரிவாக்கி 86%-இலிருந்து 90%-க்கு அவர்களின் முடிவை மேம்படுத்தினர். இதனுடன் தொடர்புடைய ஆய்வில் அவர்கள் தரப்பட்டுள்ள ஒரு கருத்தாடலுக்கு மயக்கமுள்ள சொற்கள் மிகுந்த சாத்தியத்துடன் ஒரு தனிப் பொருண்மையில் பயன்படுத்தப்படுகின்றன ("ஒரு கருத்தாடலுக்கு ஒரு பொருண்மை") என்று கூறுகின்றனர் (Gale, Church, and Yarowsky 1992c). லீகாக், சொடொரொ மற்றும் மில்லர் (Leacock, Chodorow, and Miller) என்போர் அவர்கள் செயல்பாட்டில் தலைப்புச் சூழலையும் வட்டாரச் சூழலையும் இணைத்து இவ்வலியுறுத்தலை எதிர்த்துரைத்தனர்; இது ஒரு பனுவலில் பல்பொருண்மைச் சொற்களைக் (polysemous words) கடந்து உறுதியான முடிவுகளைப் பெறத் தலைப்புச் சூழலும் வட்டாரச் சூழலும் தேவை எனக் காட்டுகிறது (பார்க்க

Towell and Voorhees 1998). யரோவ்ஸ்கியின் (Yarowsky 1993) ஆய்வு பெயர்களையும் வினைகளையும் பெயரடைகளையும் பொருண்மை மயக்கநீக்கம் செய்ய ஒரு பெரிய சாளரத்திற்குள் உள்ள தகவல்களைப் பயன்படுத்தவியலும் என்றாலும் இலக்குச் சொல்லிலிருந்து உள்ள தூரத்தைப் பொறுத்துச் சாளரத்தின் அளவு குறிப்பிடத்தக்க விதத்தில் குறைகிறது என்று காட்டுகின்றது. இது வட்டார மற்றும் தலைப்புச் சூழல்கள் பொருண்மை மயக்கநீக்கத்திற்குத் தேவை என்ற வலியுறுத்தலுக்கு ஆதரவு தருகின்றது; மேலும் வேறுபட்ட பொருண்மைமயக்கநீக்க நெறிமுறைகள் வேறுபட்ட சொல்வகைகளுக்குப் பொருத்தமாக அமையும் என்ற கருத்துச் சாயலை ஏறுமுகமாக ஏற்றுக்கொள்வதைச் சுட்டிக்காட்டுகின்றது.

ஆய்வுக்கு உட்படுத்தப்பட்டுள்ள பனுவலைத் துணைத்தலைப்புகளாகப் பகுத்துத் தலைப்புச் சூழலைப் பயன்படுத்தும் நெறிமுறைகளை மேம்படுத்தலாம். ஒரு பனுவலைப் பகுக்கும் மிக வெளிப்படையான வழி அதைப் பகுதிகளாகப் (sections) பிரிப்பதாகும்; ஆனால் இது ஒரு முனைப்பான (ஒட்டுமொத்தமான) பகுப்புதான். துணைத் தலைப்புகள் பகுதிகளுக்குள் உருவாகும்; இது பெரும்பாலும் பல பத்திகளின் ஒருங்கிணைக்கப்பட்ட குழுக்களாக அமையும். பனுவல்களைத் தானியக்கமாக அம்மாதிரியான கூறுகளாகப் பிரிப்பது தலைப்புச் சூழலைப் பயன்படுத்தும் சொற்பொருண்மை மயக்கநீக்க நெறிமுறைகளுக்குப் பயனுள்ளதாக அமையும். தொடர்ந்துவரும் கூறுகள் (segments) அல்லது வாக்கியங்களுக்குள் சொற்கள் திரும்பத்திரும்ப வருவது கருத்தாடலின் அமைப்பின் வலுவான அடையாளமாகும் (Skorochod'ko 1972; Morris 1988; Morris and Hirst 1991). பனுவல்களைத் துணைத்தலைப்புகளாகக் கூறிடும் இவ்வுற்றுநோக்கைப் (observation) பயன்படுத்தும் நெறிமுறைகள் வெளிவரத்தொடங்கியுள்ளன (Hearst 1994, van der Eijk 1994, Richmond, Smith, and Amitay 1997).

லீகாக், சொடொரோ மற்றும் மில்லர் (Leacock, Chodorow, and Miller 1998) என்போர் நுண்சூழல்-தலைப்புச்சூழல் என்பனவற்றின் பங்களிப்பைக் கருத்தில்கொண்டு அவை ஒவ்வொன்றின் பங்களிப்பை மதிப்பிட முயல்கின்றனர். ஒரு புள்ளியியல் வகைப்படுத்திக்குப் (statistical classifier) பொருண்மையின் சுட்டிக்காட்டியாக நுண் சூழல் தலைப்புச் சூழலை விட மேம்பட்டதாகும் என்று அவர்களின் முடிவுகள் காட்டுகின்றன. இருப்பினும் அண்மைக்காலச் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகள் நுண் சூழலுக்கும் தலைப்புச் சூழலுக்கும் இடையில் வேறுபாட்டைச் செய்தாலும் இவ்வேறுபாடு பொருளுள்ளதா என்பதில் தெளிவில்லை. இவ்விரண்டையும் ஒரு கோவையில் (continuum) கிடப்பதாகக் கருதுவதுவதும்

சூழல்சார் தகவல்களின் பங்களிபையும் முக்கியத்துவத்தையும் இலக்கிலிருந்து உள்ள தூரத்தின் செயல்பாடாகக் (function) கருதுவதும் மிகவும் பயனுள்ளதாக அமையும்.

3.2.1.3 பொருட்புலம் (Domain)

சொற் பொருண்மை மயக்க நீக்கத்தில் பொருட்புலத்தின் பயன்பாடு தொடக்ககால இயந்திர மொழிபெயர்ப்புச் செயன்மைகளில் உருவாக்கப்பட்ட நுண்பொருள் விளக்கச் சொற்கோவையில் (microgolssaries) முதல் சான்றாகும். பொருட்புலம் அடிப்படையில் பொருண்மைகளை மயக்கநீக்கம் செய்யும் கருத்துச்சாயல் ஷாங்கின் இயற்கைமொழி ஆய்வுக்கு உரை அணுகுமுறை (script approach) (Schank and Abelson 1977)) போன்ற பல்வேறு செயற்கை அறிவு அடிப்படையிலான அணுகுமுறைகளில் உட்படையாகும்; இது கருத்தாடலின் பொதுத் தலைப்பால் ஊக்குவிக்கப்பட்ட சூழல் அல்லது "உரை" ("script") அடிப்படையிலான பொருண்மைகளுடன் சொற்களைப் பொருத்தும். தற்போதைய கருத்தாடல் பொருட்புலத்திற்கு ஒரு சொல்லின் பொருண்மையை மட்டும் ஊக்குவிக்கும் இவ்வணுகுமுறை, இது தனியாகப் பயன்படுத்தப்படுகையில் இதன் வரம்பெல்லையை நிரூபிக்கும்; *The lawyer stopped at the bar for a drink* என்ற புகழ்பெற்ற எடுத்துக்காட்டில் சட்டத்துடன் தொடர்புடைய உரையில் உள்ள தகவலை ஒருவர் சார்ந்தால் *bar* என்பதன் தவறான பொருண்மையானது பெறப்படும்.

கேல், சர்ச் மற்றும் யரொவ்ஸ்கி (Gale, Church, and Yarowsky 1992c) என்போரின் ஒரு கருத்தாடலுக்கு ஒரு பொருண்மை என்ற வலியுறுத்தல் வாதத்திற்குரியது. டால்க்ரென் (Dahlgren 1988) பொருட்புலம் சில சொற்களுக்குப் பொருண்மை மயக்கத்தை நீக்காது என்று உற்றுநோக்கியுள்ளார்: அவர் *hand* என்ற பெயர் 16 பொருண்மைகளைக் கொண்டுள்ளன என்றும் அவற்றில் 10 பெரும்பான்மையும் எந்தப் பனுவலிலும் தக்கவைக்கப்படுகின்றது என்றும் கூறுகின்றார். பொருட்புலத்தின் தாக்கம் பனுவலின் வகை, இலக்குச் சொல்லின் பொருண்மைகளுக்கு இடையே உள்ள உறவு (வலுவாகவோ வலுவற்றோ துருவப்படுத்தப்பட்டுள்ளது, பொதுவான அல்லது சிறப்பான பயன்பாடு போன்றவை) போன்ற காராணிகளைச் சார்ந்திருக்கின்றது (எந்த அளவுக்குப் பனுவல் நுட்பமானது போன்றவை). எடுத்துக்காட்டாக French Encyclopaedia Universalis என்பதில் *intèrêt* ("interest") என்ற சொல் INTEREST-FINACE என்ற கட்டுரையில் 62 தடவைகள் தோன்றுகின்றது; இவ்வெல்லா நேர்வுகளிலும் அது நிதிசார்ந்த (financial) பொருண்மையில் பயன்படுத்தப்படுகின்றது; இச்சொல் INTEREST-PHILOSOPHY AND HUMANITIES என்ற கட்டுரையில் 139 தடவைகள் அதன்

பொதுவான நிதியல்லாத பொருண்மையில் தோன்றுகின்றது. இருப்பினும் THIRD WORLD என்ற கட்டுரையில் intèrêt என்ற சொல் இவ்விரு பொருண்மைகளிலும் இரு தடவைகள் தோன்றுகின்றது.

3.2.2 பொருண்மை பகுப்பு

3.2.2.1 வங்கி மாதிரி

சொற்பொருள் மயக்கநீக்கதில் பெரும்பாலான ஆய்வாளர்கள் தற்போது இயந்திரத்தால் படிக்கவியலும் அகராதிகள் (machine-redable dictionaries) அல்லது சொல்வலை (wordNet) (இது OALD-இன் பொருண்மைகளைப் பயன்படுத்துகின்றது) போன்ற நிறுவப்பட்ட சொல்சார் மூலவளங்களால் தரப்பட்ட பொருண்மை வேறுபாடுகளைச் சார்ந்திருக்கின்றனர்; ஏனென்றால் அவை பரவலாகக் கிடைக்கின்றன. இவ்வாய்வுகளில் ஆதிக்கமுள்ள மாதிரி "வங்கி மாதிரியாகும்" ("bank model"); இது bank-money மற்றும் bank-riverside என்பவைகளுக்கு இடையில் தெளிவான எல்லைக்கோட்டை எல்லாப் பொருண்மை வேறுபாடுகளுக்கும் நீட்சிசெய்ய முயன்றது. இருப்பினும் இவ்வசதியான எல்லைக்கோட்டை எல்லாச் சொற்களுக்கோ அல்லது பெரும்பாலான பிற சொற்களுக்கோ பயன்படுத்த வழியில்லை. பொருண்மைகளின் கருத்துச்சாயலுக்குச் சிறிது உளவியல் மதிப்பு இருந்தாலும் (Simpson and Burgess 1988; Jorgensen 1990), அகராதியியலார்கள் பொருண்மைகள் மற்றும் பொருண்மைப் பகுப்புகள்/பிரிவுகள் பற்றி உடன்பாடு இல்லாமை பற்றி நன்கு அறிவார்கள் (Malakhovski 1987, Robins 1987, Ayto 1983, Stock 1983). பொருண்மைப் பகுப்புச் சிக்கல் பழங்காலத்திலிருந்தே வாதிடும் பொருளாக இருந்தது: அரிஸ்டாட்டில் 350-களில் தன்னுடைய Topics என்பதில் ஒரு பகுதியை இதற்கு என்று ஒதுக்கி வைத்தார். அதிலிருந்து தத்துவயியலார்களும் மொழியியலார்களும் இத்தலைப்பு பற்றி விரிவாகத் தொடர்ந்து வதிட்டுவருகின்றனர் (பார்க்க Quine 1960, Asprejan 1974, Lyons 1977, Weinrich 1980, Cruse 1986); ஆனால் 2000 ஆண்டுகளாகத் தீர்மானம் இல்லாதிருப்பது கவனத்திற்குரியது.

3.2.2.2 நுணுக்கம் (Granularity)

சொற்பொருண்மை மயக்கநீக்கத்தின் மிக முக்கியமான சிக்கல் பொருண்மை நுணுக்கத்தின் (Granularity) பொருத்தமான அளபை நிர்ணயிப்பதாகும். பல படைப்பாளிகள் (எடுத்துக்காட்டாக Slator and Wilks 1987) அகராதியில் நாம் பார்க்கும் பொருண்மைப் பிரிவுகள் இயற்கை மொழி ஆய்வுச் செயன்மையின் தேவைக்கு அதிகமான நுணுக்கமானதாகும் என்று

கூறியுள்ளனர். கூடுதல் நுணுக்கமான பொருண்மை வேறுபாடுகள் தானியக்கச் சொற்பொருண்மை மயக்கநீக்கத்திற்கு நடைமுறைச் சிக்கல்களை உருவாக்கும்: அவர்கள் சிறப்புத்தன்மையான ஒன்றுசேர்ப்பு விளைவுகளை (combinatorial effects) அறிமுகம் செய்கின்றனர் [எடுத்துக்காட்டாக ஸ்லாடர் மற்றும் வில்க்ஸ் (Slator and Wilks 1987) என்போர் நடுத்தர அளவுள்ள LDOCE-ஐப் பயன்படுத்தி *There is a huge envelope of air around the surface of the earth* என்பதற்கு 284,592 வேறுபட்ட ஆற்றல்மிக்க ஒருங்கிணைந்த பொருண்மை ஒதுக்கீடுகள் (combined sense assignments) உள்ளன என்று கண்டுள்ளனர்]; அவர்கள் திறமையான அகரதியியலார்களுக்கே மிகச் சிக்கலாக இருக்கும் பொருண்மை விருப்பத்தேர்வுகளை வேண்டினர்; மேலும் அவர்கள் நிறைவேறாத விகிதங்களில் கண்காணிக்கப்பட்ட நெறிமுறைகளுக்குத் தேவையான தரவின் அளவை அதிகரித்தனர். இத்துடன் பல அகராதிகளில் செய்யப்பட்டுள்ள பொருண்மை வேற்றுமைகள் (sense distinctions) சில வேளைகளில் படிப்பவர்களின் (வேறுபடுத்தும்) திறனுக்கு அப்பாற்பட்டவையாகும். கில்காரிஃப் (Kilgarriff 1992, 1993) தமது மிகவும் அறியப்பட்ட ஓர் ஆய்வில் LDOCE-இல் உள்ள ஒரு தனித்தன்மைவாய்ந்த பொருண்மைக்குப் (unique sense) பல சொற்களை ஒதுக்குவது படிப்பவர்களுக்கு இயலாததாகும் என்று காட்டுகின்றார். இதைப் புரிந்து கொண்டு டொலான் (Dolan 1994) முனைப்பான பொருண்மை வேற்றுமைகளை உருவாக்க அகராதிப் பொருண்மைகளை இணைத்து அவற்றை மயக்கம் செய்ய ஒரு நெறிமுறையை முன்மொழிந்தார். பிறர் ரோஜஸ்ட்டின் சொற்களஞ்சியம் போன்ற முனைப்பான பொருண்மைப் பகுப்புகளைப் பயன்படுத்தினர்; இருப்பினும் ஒரு தனித்துவமான பொருண்மையை ஒதுக்குவதோ அல்லது விருப்புத்தேர்வுக்குள் ஒரு பொருத்தமான ஒன்றைக் கண்டுபிடிப்பதோ பெரும்பாலும் கடினமானதாகும் (எடுத்துக்காட்டாகப் பார்க்க Yarowsky 1992). சென் மற்றும் சாங் (Chen and Chang (1998) என்போர் ஒரு அகராதியில் (LDOCE) அர்த்தங்களை ஒன்றுசேர்த்து அவற்றை ஒரு சொற்களஞ்சியத்தின் (LLOCE) வகைபாடுகளுடன் (categories) இணைக்க ஒரு வழிமுறை வரைவை (algorithm) முன்மொழிந்தனர்.

அகராதி அர்த்தங்களை ஒன்றுசேர்ப்பது சிக்கலைத் தீக்காது. முதலாவதாக நுண்மையின் தேவையின் அளபு மேற்கொண்ட செயல்பாடைப் பொறுத்தது. பல்பொருள் ஒருமொழி (ஒப்புருமொழி) வேற்றுமை மட்டுமே பேச்சு உருவாக்கத்திற்கும் பனுவலில் ஒலியழுத்தங்களின் தக்கவைப்புக்கும் தேவையாகும்; அதேசமயம் இயந்திர மொழிபெயர்ப்பு போன்ற செயல்பாடுகள்

நுண்ணிய பொருண்மை வேற்றுமைகளை வேண்டும்; சில நேர்வுகளில் ஒருமொழிய அகராதி (monolingual dictionary) தருவதைவிட நுண்ணிய வேற்றுமைகள் தேவை (எடுத்துக்காட்டாகப் பார்க்க ten Hacken 1990). எடுத்துக்காட்டாக *river* என்ற ஆங்கிலச் சொல் பிரஞ்சு மொழியில் ஆறு கடலை நோக்கி ஒழுகும் போது *fleuve* என்றும் அவ்வாறில்லாவிடில் *rivière* என்றும் மொழிபெயர்ப்பு செய்யப்படுகின்றது. இருப்பினும் தரப்பட்ட செயல்பாட்டுக்கும் நுணுக்கத்தின் தேவையின் அளபுக்கும் ஒரு கட்டுப்பாடான பொருத்தம் இல்லை. எடுத்துக்காட்டாக, முன்னர் பார்த்தபடி, *mouse* என்ற சொல் இரு வேறுபட்ட பொருண்மைகளைக் (*animal, device*) கொண்டிருந்தாலும் பிரஞ்சு மொழியில் இரண்டு நேர்விலும் *souris* என்றுதான் மொழிபெயர்க்கப்படுகின்றது. மாறாகத் தகவல் மீட்பிற்கு *mouse* என்பதன் இரு பொருண்மைகளுக்கு இடையே உள்ள வேற்றுமை முக்கியமாகும்; அதே சமயம் *fleuve* என்ற அர்த்தத்திலுள்ள *river* என்பதற்கும் *rivière* என்ற அர்த்ததிலுள்ள *river* என்பதற்கும் வேறுபாடு காணவேண்டியதன் காரணத்தைப் பற்றிச் சிந்திப்பது கடினமாகும். இரண்டவதாக மற்றும் பொதுவாக எப்போது பொருண்மைகள் சேர்க்கப்படவேண்டும் அல்லது பிரிக்கப்படவேண்டும் என்பதில் தெளிவில்லை. அகராதியலார்களும் உடன்படுவதில்லை: பில்மோரும் அட்கின்ஸ்சும் (Fillmore and Atkins 1991) *risk* என்ற சொல்லின் மூன்று பொருண்மைகளை அடையாளம் காண்கின்றனர்; ஆனால் பெரும்பாலான அகராதிகள் குறைந்தது அவைகளில் ஒன்றைக்கூட பட்டியலிடுவதில்லை என்று கண்டுகொண்டனர். பல நேர்வுகளில் அர்த்தமானது அர்த்ததின் நிழல்கள் விழும் ஒரு கோவையாகக் கருதப்படுகின்றது (எடுத்துக்காட்டாக Cruse 1986); மேலும் பொருண்மைகள் ஒன்றுசேரும் அல்லது பிரியும் இடங்கள் விரைந்து மாற்றமுறும்.

3.2.2.3 பொருண்மைகளா பயன்பாடுகளா (senses or usages?)

சொற்கள் குறிப்பிட்ட பொருண்மைகளுடனும் கருத்துருக்களுடனும் பொருந்தும் என்ற அரிஸ்ட்டாடலியன் கருத்து ஸசூர் மற்றும் பிறரால் (Meillet 1926, Hjemslev 1953, Martinet 1960) வெளிப்படுத்தப்பட்டது. எடுத்துக்காட்டாக ஆன்டொனி மில்லெடைப் பொறுத்தவரையில் ஒரு சொல்லின் பொருண்மை அதன் சராசரி மொழிசார் பயன்பாடுகளால் மட்டும்தான் வரையறை விளக்கம் செய்யப்படும். விட்ஜென்ஸ்டைன் (Wittgenstein 1953) தனது Philosophische Utersuchungen-இல் பொருண்மைகளே இல்லை பயன்படுகள் மட்டும் தான் உள்ளன என்பதை வலியுறுத்தி இதுபோன்ற நிலைபாட்டை எடுக்கின்றார்.

> “நாம் “அர்த்தம்” என்ற சொல்லைப் பயன்படுத்தும் நேர்வுகளின் பெரிய வகுப்புக்கு (எல்லாவற்றிற்கும் இல்லாவிட்டாலும்) அதை இவ்வாறு வரையறை விளக்கம் செய்யலாம்: ஒரு சொல்லின் அர்த்தம் மொழியில் அதன் பயன்பாடாகும்.” (Sect. 43 1953).

மிக அண்மைக்கால அர்த்தக் கோட்பாடுகளில் இதுபோன்ற கருத்துக்கள் வெளிப்படையாகும்; எடுத்துக்காட்டாக புளூம்பீல்டு (Bloomfield 1933), ஹாரிஸ் (Harris 1954) என்பவர்களுக்கு அர்த்தம் என்பது வினியோகத்தின் செயல்பாடாகும்; மேலும் பார்வைஸ் மற்றும் பெரி (Barwise and Perry 1953) என்போரின் சூழல் பொருண்மையியலில் (situation semantics) ஒரு சொல்லின் பொருண்மை அல்லது பொருண்மைகள் கருத்தாடலில் அது சீராக நடத்தைசெய்யும் பங்கேற்பின் ஒரு கருத்துப்பொருளாகப் (abstraction) பார்க்கப்படுகின்றது.

COBUILD திட்டம் (Sinclair 1987) ஒரு தரவுத்தொகுதியில் மேற்கோள்களின் கொத்துக்களின் (கூட்டங்களின்) அடிப்படையில் பொருண்மைப் பிரிவுகளை உருவாக்கித் தற்காலப் பயன்பாட்டில் அகராதிப் பொருண்மைகளை நிறுவ முயற்சிசெய்து அர்த்தத்தின் இப்பார்வையை ஏற்றுக்கொள்கின்றது. அட்கின்ஸ் (Atkins 1987) மற்றும் கில்கரிஃப் (Kilgarrif) என்போர் ஹாரிஸ்சின் (Harris 1954) பார்வையை உட்படையாக ஏற்றுக்கொள்கின்றனர்; இதன்படி ஒவ்வொரு பொருண்மை வேறுபாடும் (sense distinction) ஒரு தனித்தன்மையான சூழலில் பிரதிபலிக்கின்றது. இதே போன்ற பார்வை முன்னர் மேற்கோள்காட்டப்பட்ட வகுப்பு-அடிப்படியிலான நெறிமுறைகளை உள்ளுறை செய்கின்றது (Brown et al. 1992; Pereira and Tishby 1992; Pereira, Tishby and Lee 1993). ஸ்குட்சே (Scutze 1998) இத்தனிப் போக்கைத் தொடர்கின்றார்; மேலும் பொருண்மை வேறுபாட்டின் சிக்கலை முற்றிலும் தவிர்க்கும் ஒரு உபாயத்தை முன்மொழிகின்றார்: அவர் முன்நிறுவப்பட்ட பொருண்மைப் பட்டியலைச் சாராமல் ஒரு தரவுத்தொகுதியிலிருந்து பொருண்மைக் கொத்துக்களை (sense clusters) உருவாக்குகின்றார்.

3.2.2.4 பட்டியலிடுதலா அல்லது ஆக்குதலா? (Enumeration or generation?)

ஆக்கமுறை அகராதிகளின் உருவாக்கம் (Pustejovsky 1995) இன்றுவரையுள்ள எல்லா சொற்பொருண்மை மயக்கநீக்கச் செயன்மைகளிலிருந்தும் மிகவும் வேறுபட்ட சொற்பொருண்மையின் பார்வையைத் தருகின்றது. பட்டியலிடும் அணுகுமுறை (enumerative approach) சூழலிருந்து சுதந்திரமாக இருக்கும் நிலைபெறான பொருண்மைகளின் குழுமம் என்ற

விதிதரு முறையில் அமைந்த ஒன்றை ஊகமாய்க்கொள்கின்றது; இது அடிப்படையில் அரிஸ்டாட்டிலியன் பார்வையாகும். சூழல் கருத்தில் கொள்ளப்படாதவரை பொருண்மை வகுத்தொதுக்குதல்கள் (sense assignments) குறைத் திட்டவட்டமானவை (underspecified) என்று ஊகம் செய்துகொண்டு ஆக்கமுறை அணுகுமுறையானது பொருண்மையின் கருத்தாடல் சார்ந்த உருப்படுத்தத்தை உருவாக்குகின்றது.

சொற் பொருண்மை மயக்க நீக்கத்திற்குத் தேவையான மற்றும் பொருத்தமான பொருண்மைக் குழுமங்களை நிர்ணயிக்கும் கடினங்களைக் கருத்தில் கொண்டு, சொற்பொருண்மை மயக்கநீக்க ஆய்வில் ஆக்கமுறை பார்வையின் திறனுக்குச் சிறிதுதான் கவனம் செலுத்தப்பட்டுள்ளது என்பது ஆச்சரியத்திற்குரியதாகும். பெரிய மற்றும் முழுவதுமான ஆக்கமுறை அகராதிகள் கிடைக்கும் போது தான் பொருண்மை வகுத்தொதுக்குதலுக்கு இவ்வணுகுமுறையை ஆய்வது தகுதியுள்ளதாக அமையும்.

3.3. முடிவுரை

தானியக்க மொழிப் பகுப்பாய்வு (automatic language processing) இருக்கிறது வரை சொற்பொருண்மை மயக்கநீக்க ஆய்வுகள் ஒரு வரலாறாக வந்து கொண்டிருக்கும். பின்னோக்கிப் பார்த்தால் பெரும்பாலான சிக்கல்களும் அவற்றிற்கான அணுகுமுறைகளும் தொடக்ககாலத்திலேயே புரிந்துகொள்ளப்பட்டுள்ளன என்பது குறிப்பிடத்தக்கதாகும். சொற்பொருண்மை மயக்கநீக்கத்தின் மீதான தொடக்ககால ஆய்வுகள் ஒத்தறி அடிபடையில் மறைக்கப்பட்ட/மறக்கப்பட்ட நூல்களாகவும் கட்டுரைகளாகவும் பல ஆய்வுக்களங்களிலும் துறைகளிலும் வெளிவந்ததன் காரணமாகத் தற்போதைய ஆய்வாளர்களும் படைப்பாளிகளும் அவற்றைப் பற்றி அறியாதிருக்கின்றனர் என்றுகூற இயலும். கடந்த 50 வருடங்களில் சொற்பொருண்மை மயக்கநீக்க ஆய்வில் மிகக்குறைவான முன்னேற்றமே எய்தப்பட்டுள்ளது என்பது ஆச்சரியப்படத்தக்கதாகும். அண்மைக்கால ஆய்வுப்பணிகள் 90% அல்லது அதற்கு மேற்பட்ட விளைவுகளை/முடிவுகளைக் குறிப்பிட்டாலும் இவ்வாய்வுகள் எடுத்துக்காட்டாகச் சில சொற்களையே, பெரும்பாலும் பெயர்ச் சொற்களையே உட்படுத்தியுள்ளன அல்லது எடுத்தாண்டுள்ளன; அதிலும் பரந்த அர்த்த வேறுபாடுகளையே செய்துள்ளன.

சொற்பொருள் மயக்கநீக்க ஆய்வுகள் ஒரு முழு சுற்று சுற்றி தொடக்ககாலத்தில் சொற்பொருண்மை மயக்கநீக்கச் சிக்கல்களுக்குத் தீர்வு காணப் பயன்படுத்தப்பட்ட அனுபவாத அணுகு முறைகளுக்கும் தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகளுக்கும் மீண்டும்

வந்துள்ளன. 1990களில் போதுமான அளவிலான மூலவளங்கள் மற்றும் மேம்படுத்தப்பட்ட புள்ளியியல் அணுகுமுறைகள் காரணமாகத் தொடக்ககால முடிவுகளிலும் விளைவுகளிலும் முன்னேற்றம் ஏற்பட்டுள்ளன; இருப்பினும் நாம் தற்போதைய சட்டகத்தில் எவ்வளவு சாதிக்க இயலுமோ அதன் எல்லையை அடைந்துவிட்டோம். இதன் காரணமாக நாம் சொற்பொருண்மை மயக்கநீக்க ஆய்வுகளின் தற்போதைய நிலையைப் பற்றி மதிப்பிடுவது அவசியமாகும். அத்தகைய முயற்சி இங்கு மேற்கொள்ளப்பட்டுள்ளது.

இயல் 4

சொற்பொருண்மை மயக்கநீக்கத்தின் செயல்பாடுகளும் முன்னேற்றங்களும்

4.0 முன்னுரை

சொற்பொருண்மை மயக்கநீக்கம் என்பது ஒரு கணினிசார் வழி சூழலில் சொற்களின் பொருளைக் கண்டுகொள்ளும் ஒரு திறனாகும். சொற்பொருண்மை மயக்கநீக்கம் ஒரு செயற்கை அறிவு-முழுமைச் சிக்கலாகக் (AI-complete problem) கருதப்படுகின்றது; அதாவது, செயற்கை அறிவில் மிகக் கடினமான சிக்கல்களுக்குச் சமமான செயல்பாடாகும். இங்கே சொற்களின் மயக்கத்தை நீக்குவதற்கான செயலாக்கங்களும் இச்செயலின் விளக்கமும் தரப்படும். இங்கு கண்காணிக்கப்பட்ட, கண்காணிக்கப்படாத மற்றும் அறிவு அடிப்படையிலான அணுகுமுறைகள் சுருக்கமாகக் கூறப்படும். சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் மதிப்பீடு பல்வேறுபட்ட மயக்கநீக்கச் செயல்படுகளில் பங்கெடுக்கும் ஒழுங்குமுறைகளின் புறவயமான மதிப்பீட்டை நோக்கமாகக்கொண்டு Senseval/Semeval நடவடிக்கைகளின் சூழலில் விவாதிக்கப்படும். இறுதியாயாக பயன்பாடுகள், வெளிப்படையான சிக்கல்கள் மற்றும் எதிர்காலத் திசைகள் விவாதிக்கப்படும்.

4.0.1 நோக்கம்

மனித மொழி மயக்கமுள்ளது; பல சொற்கள் அவை வரும் சூழலைப் பொறுத்து பல வழிகளில் பொருள்கோள் செய்யவியலும். எடுத்துக்காட்டாக, பின்வரும் வாக்கியங்களை எடுத்துகொள்ளவும்:

அ. I can hear bass sounds.

ஆ. They like grilled bass.

இந்த இரு வாக்கியங்களிலும் *bass* என்ற சொல்லின் நேர்வுகள் வேறுபட்ட பொருள்களைக் குறிப்பிடுகின்றது: முறையே குறைந்த நிகழ்வெண் மற்றும் ஒருவகை மீன்.

துரதிருஷ்டவசமாக, சூழலில் சொல் பெறும் குறிப்பிட்ட பொருளின் கண்டுபிடிப்பு மேலோட்டமாகத்தான் எளிதாகத் தோன்றுகின்றது. பொரும்பாலன வேளைகளில் மனிதர்கள் மொழியிலுள்ள மயக்கங்களைப் பற்றிச் சிந்திப்பதே இல்லை; அதே சமயம் இயந்திரங்கள் (கணினிகள்) அமைப்பாக்கம் செய்யப்படாத பனுவல் தகவல்களைப் பகுப்பாய்வு செய்யவேண்டும்; மற்றும் அவற்றை உள்ளுறும் பொருண்மையை நிர்ணயிப்பதற்காகப்

பகுத்தாயப்படவேண்டிய தரவு அமைப்பாக மாற்ற வேண்டும். சூழலில் சொற்களின் பொருண்மைகளின் கணினிசார் கண்டுபிடிப்பு சொற்பொருண்மை மயக்கநீக்கம் எனப்படும். எடுத்துக்காட்டாக, மயக்கநீக்கத்தின் விளைவாக, முன்னர் கூறப்பட்ட ஆ என்ற வாக்கியம் பின்வருமாறு சீர்மையாக அர்த்தம் அடையாளப்படுத்தப்பட (sense-tag) வேண்டும்: “They like/$_{ENJOY}$ grilled/$_{COOKED}$ bass/$_{FISH}$.”

சொற்பொருண்மை மயக்கநீக்கம் ஒரு செயற்கை அறிவு-முழுமைச் சிக்கலாக (AI-complete problem) விளக்கப்பட்டுள்ளது (Mallery 1988); அதாவது, செயற்கை அறிவின் மையச் சிக்கல்களை நீக்குவதற்குச் சமமாகன கடினமான கலவைக் கோட்பாடில் பெயர்த்தொடர்-முழுமைக்கு (NP-completeness) ஒப்புமையால் நிகராகும்; எடுத்துக்காட்டாக, டர்னிங் பரிசோதனை (Turning Test [Turning 1950]). இதன் ஒப்புக்கொள்ளப்பட்ட கடினம் ஒரு காரணத்தால் உருவானதல்ல; பல்வேறுபட்ட காரணிகளிலிருந்து உருவானதாகும்.

முதலாவது, சொல் மயக்க நீக்கத்தின் உருப்படுத்ததின் அணுகுமுறை (ஒரு முற்று குழும அர்த்தங்களின் கணக்கெடுப்பிலிருந்து புதிய அர்த்தங்களின் விதி அடிப்படையிலான உருவாக்கம் வரையிலான வியாபிப்பு), அர்த்தத் தெரிவடைவுகளின் நுண்மை (granularity of sense of inventories) (நுண்மையான வேறுபாடுகளிலிருந்து உள்ளடங்கு மொழிகளுக்கு), பொருட்புலச் சார்புப் பனுவல்கள் (domain-oriented texts) மற்றும் அதற்கு எதிரிடையான பனுவல்களின் கட்டுப்பாடற்ற இயல்பு (unrestricted nature of texts), பொருண்மை மயக்கநீக்கத்திற்கான இலக்குச் சொற்களின் குழுமம் (ஒரு வாக்கியத்திற்கு ஒரு இலக்குச் சொல் மற்றும் இதற்கு எதிரிடையான “எல்லாச் சொற்கள்” என்ற பின்னணியமைப்பு) போன்ற அடிப்படைக் கேள்விகளின் வேறுபட்ட முறையாக்கங்களுக்குச் செயற்பாடு விளக்கம் தருகின்றது.

இரண்டாவது சொற்பொருண்மை மயக்கநீக்கம் அறிவில் கூடுதலாகச் சார்ந்திருக்கின்றது. உண்மையில், எந்தச் சொற்பொருண்மை மயக்கநீக்கத்தின் விரிவற்ற செயல்முறையையும் பின்வருமாறு சுருக்கவுரை செய்யலாம்: சொற்கள் தரப்படுகையில் (எ.கா. ஒரு வாக்கியம் அல்லது சொற்களின் ஒரு பை), சூழலில் சொற்களுடன் மிகப் பொருத்தமான அர்த்தங்களைச் சேர்ப்பதற்கு ஒன்றோ அதற்குக் கூடுதலோ மூலங்களைப் பயன்படுத்தும் ஒரு தொழில்நுட்பம் பயன்படுத்தப்படுகின்றது. அறிவு மூலங்கள் புலக்குறிப்பு செய்யப்படாத அல்லது சொல் அர்த்தங்களுக்கு அடையாளப்படுத்தப்பட்ட பனுவல்களின் தரவுத்தொகுதிகளிலிருந்து (அதாவது

சேகரிப்புகள்) இயந்திரத்தால் படிக்கவியலும் அகராதிகள், பொருண்மை வலையமைப்புகள் போன்றன வரை கருதத்தக்க விதத்தில் வேறுபடவியலும். அறிவு இல்லாமல் மனிதர்களுக்கும் இயந்திரங்களுக்கும் பொருண்மையை (எடுத்துக்காட்டாக முந்தைய வாக்கியங்களில்) அடையாளம்காண இயலாது.

துரதிருஷ்ட வசமாக, அறிவு மூலவளங்களின் மனித உருவாக்கம் விலையுயர்ந்த மற்றும் காலவிரயமாகும் முயற்சியாகும்; இம்முயற்சி சொற்பொருண்மை மயக்க நீக்கத்தின் விவரக்குறிப்பு மாறும் ஒவ்வொரு தடவையும் திரும்பச் செய்யப்படவேண்டும் (எ.கா. புதிய பொருட்புலங்கள், வேறுபட்ட மொழிகள் மற்றும் அர்த்தங்களின் தெரிவடைவுகள் இவற்றின் இருப்பில்). இந்த அடிப்படையான சிக்கல் தான் சொற்பொருண்மை மயக்க நீக்கக் களத்தில் பரவியிருக்கின்றது; இது அறிவுப் பேறு நெருக்கடி (knowledge acquisition bottleneck) என்று அழைக்கப்படுகின்றது (Gale et al. 1992b).

சொற்பொருண்மை மயக்கநீக்கத்தின் கடினத்திற்கு, உண்மை-உலகச் செயல்பாடுகளுக்கு இதன் குறைவான பயன்பாடு சான்றாக அமைகின்றது. தகவல் தொழில்நுட்பத்தின் (information technology (IT)) பல இடங்களில் விரைவான முன்னேற்றத்துடன் கூடிய இணையச் சமூகத்தின் அதிக அளவிலான வளர்ச்சியானது ஆவணக்கிடங்குகள் (document warehouses), வலைப் பக்கங்கள் (web pages), அறிவியல் கட்டுரைகளின் சேகரிப்புகள், பிளாக் தரவுத்தொகுதி (blog corpora) போன்ற அமைப்பாக்கம் செய்யப்படாத தரவின் மிக அதிகமான அளவிலான உற்பத்திக்கு வழிவகுத்துள்ளது. இதன் விளைவாகத் தானியக்க நெறிமுறைகளின் வழி இந்தத் தகவலின் பெருந்திரளைக் கையாளுவதற்கு அதிகரிக்கும் தூண்டுதல் உள்ளது. பனுவல் அகழ்வுக்கும் (text mining) தகவல் மீட்புக்கும் (information retrieval) பயன்படுத்தப்பட்ட மரபான உபாயங்கள் மிக அதிக அளவிலான தகவலின் சேமிப்புகளில் பயன்படுத்தப்படுகையில் அவற்றின் வரம்பெல்லையைக் காட்டுகின்றன. உண்மையில் பெரும்பான்மையும் பனுவலின் சொல்-தொடரியல் ஆய்வு (lexicosyntactic analysis) அடிப்படையிலான இந்த அணுகுமுறைகள் சொற்களின் புறத் தோற்றத்தைத் தாண்டி போவதில்லை; இதன் விளைவாக வேறுபட்ட சொற்களினால் உருவாக்கப்பட்ட தகுதியுள்ள தகவலை அடையாளம் காணவும் பயன்படுத்துபவரின் தேவைக்குப் பொருந்தாத ஆவணங்களைக் கைவிடவும் தவறுகின்றன. பனுவல் பொருண்மைமயக்க நீக்கம் (text disambiguation) மிகப் பெரிய அளவிலான தரவைக்

கையாளுவதில் ஒரு முக்கிய வழியைத் தரவியலும்; இதன்படி பொருண்மை வலை (Semantic Web) என்று அழைக்கப்படுகின்ற ஒன்றுக்கு அடிப்படையான பங்களிப்பு செய்கின்றது. பொருண்மை வலை தற்போதைய வலை (Web) என்பதன் நீட்சியாகும்; இதில் தகவல் நன்றாக வறையறுக்கப்பட்ட பொருண்மையில் தரப்பட்டுள்ளது; இது கணினிகளையும் மக்களையும் ஒற்றுமையாக வேலை செய்யச் சாத்தியப்படுத்தும் (Berners-Lee et al. 2001, page 2).

நாம் இயந்திர மொழிபெயர்ப்பின் சிக்கலைக் கூறும் போது சொற்பொருண்மை மயக்கநீக்கத்தின் திறன் தெளிவாகும்: *penna* என்ற இத்தாலியச் சொல் சூழல் அடிப்படையில் ஆங்கிலத்தில் *feather*, *pen* அல்லது *author* என்று மொழிபெயர்க்கப்படலாம். பனுவலின் தானியக்க மொழிபெயர்ப்பில் (automated translation) சொற்பொருண்மை மயக்கநீக்கம் முக்கியமான பங்களிப்பு செய்யும் இம்மாதிரியான நேர்வுகள் ஆயிரக்கணக்கில் உள்ளன; இது சொற்பொருண்மை மயக்கநீக்கத்தின் வரலாற்றுப் பயன்பாடாகும்.

சொற்பொருள் மயக்கநீக்கம் எடுத்துகாட்டான ஒரு இடைப்பட்டச் செயல்பாடாக ஒழுங்கமைக்கப்படுள்ளது; அதாவது ஒரு தனியாக நிற்கும் தொகுதியாகவோ (stand-alone module) ஒரு பயன்பாட்டில் பொருத்தமாக ஒருங்கிணைக்கப்படவோ செய்யப்பட்டுள்ளது (இதன்படி உள்ளுறையாக பொருண்மை மயக்கநீக்கத்தைச் செயல்படுத்தும்). இருப்பினும், உண்மை உலகில் சொற்பொருண்மை மயக்கத்தின் பயன்பாடுகளின் வெற்றி வெளிப்படுத்தப்பட வேண்டிய நிலையிலேயே உள்ளது. இத்தலைப்பில் வேறுபட்ட செயல்பாடுகளும் (ஆய்வுகளும்) கோரிக்கைகளும் (proposals) வெளியிடப்பட்டிருந்தாலும், சொற்பொருண்மை மயக்கநீக்கத்தின் பயன்பாடு அடிப்படையிலான மதிப்பீடு ஒரு திறந்த ஆய்வுக் களமாகும்.

சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் சமீபகால ஒப்பீட்டு மதிப்பீடுகளின் முடிவுகள் (பெரும்பாலும் சொற்பொருண்மை மயக்கநீக்கத்தின் தனிநிலை மதிப்பீட்டைக் கருத்திற்கொண்டவை) நுணுக்கமான பொருண்மை வேறுபாடுகள் நடைமுறைபடுத்தப்படுகையில் செயல்திறன் மற்றும் பொதுமையாக்கும் திறன் அடிப்படையில் பெரும்பாலான பொருண்மை மயக்கநீக்க நெறிமுறைகள் மரபுரிமையாகவே வரம்பைக் கொண்டிருப்பதைக் காட்டுகின்றன. மாறாக விரிந்த-உள்ளடக்கப்பரப்பு, செழுமையான சொல் அறிவு மூலவளங்கள், மற்றும் மிக அதிக அளவிலான நுண்மையாக்கப்படாத பொருண்மை தெரிவடைகளின் உருவாக்கம் என்பன பொருண்மை மயக்கநீக்க அணுகுமுறைகளுக்குப் புதிய சந்தர்ப்பங்களைத் தருவதாகத்

தெரிகின்றது; குறிப்பாக மனித-மொழி தொழில்நுட்பத்தின் களத்தில் பொருண்மை அடிப்படையில் சாத்தியமான பயன்பாடுகளை நோக்கமாகக்கொள்கையில் புதிய சந்தர்ப்பங்களைத் தருகின்றது.

4.0.2 சுருக்கமாக வரலாறு

இயற்கை மொழி ஆய்வு (Natural Language Processing (NLP)). உண்மையில், 1940களிலேயே (Weaver 1949) சொற்பொருண்மை மயக்கநீக்கம் இயந்திர மொழிபெயர்ப்பில் அடிப்படைச் செயல்பாடாகக் கருதப்பட்டது. அச்சமயத்தில் ஆய்வாளர்கள் தங்கள் மனதில் சொற்பொருண்மை மயக்கநீக்கத்தின் ஆக்கக்கூறுகளை (பகுதிகளை) எண்ணிவைத்திருந்தனர்: ஒரு இலக்கு மொழி சொல் வரும் சூழல், சொற்கள் மற்றும் பொருண்மைகளைப் பற்றிய புள்ளியியல் தகவல்கள், அறிவு மூலவளம் போன்றன. மிக விரைவில் சொற்பொருண்மை மயக்கநீக்கம் ஒரு கடினமான சிக்கல் என்பது தெளிவானது; மேலும் கணினியாக்கத்திற்குத் தேவையான எல்லைக்குட்பட்ட வழிகள் தான் உள்ளது என்பதும் தெளிவானது. 1960களில் இயந்திர மொழிபெயர்ப்பின் முன்னேற்றதிற்கு சொற்பொருண்மை மயக்கநீக்கத்தின் கடினம்தான் முக்கியமான தடைகளில் ஒன்றாக ஒத்துக்கொள்ளப்பட்டது (Bar-Hillel 1960). 1970களில் மொழிப் புரிந்துகொள்கையை இலக்காகக் கொண்டு செயற்கை அறிவு அணுகுமுறைகளின் துணையுடன் சொற்பொருண்மை மயக்கநீக்கத்தின் சிக்கல் எதிர்கொள்ளப்பட்டது. இருப்பினும் இயந்திரம் படிக்கவியலும் அறிவின் (machine-readable knowledge) குறைவு காரணமாக இதன் விளைவுகளைப் பொதுமையாக்கம் செய்வது கடினமாக இருந்தது. இந்த அடிப்படையில், அறிவுப் பிரித்தெடுத்தலுக்கு (knowledge extraction) (Wilks et al. 1990) தானியக்க நெறிமுறைகளைச் சாத்தியமாக்கும் மிக அதிக அளவிலான சொற்சார் மூலவளங்களின் வெளியீடு காரணமாக 1980களில் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடுகள் ஒரு திருப்புமுனையை அடைந்தது. 1990கள் புள்ளியியல் நெறிமுறைகளின் மிக அதிக அளவிலான பயன்பாட்டையும் இன்று வரை சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் பருவ அடிப்படியிலான மதிப்பீட்டு நடவடிக்கைகளுக்கும் கொண்டுசென்றது.

4.0.3 சுருக்கம்

இவ்வியலில் சொற்பொருண்மை மயக்கநீக்கச் செயல்பாடு முதலில் முறைபடுத்தப்படும்; பின்னர் முக்கியமான அணுகுமுறைகள் விளக்கப்படும். அடுத்தபடியாகச் சொற்பொருண்மை

மயக்கநீக்கத்தின் மதிப்பீடும் அதைத் தொடர்ந்து உண்மை-உலகப் பயன்பாட்டுகளில் அதன் திறனும் விளக்கப்படும். பின்னர் திறந்த சிக்கல்களும் எதிர்காலத் திசைகளும் ஆயப்படும். இறுதியில் முடிவுரை தரப்படும்.

4.1 செயல்பாட்டு விளக்கம்

ஒரு சூழலில் ஒரு சொல்லின் எந்த அர்த்தம் தூண்டப்படுகின்றது என்பதைக் கணினி மூலம் தீர்மானிக்கும் திறன் தான் சொற்பொருண்மை மயக்கநீக்கம் ஆகும். சொற்பொருண்மை மயக்கநீக்கம் ஒன்றோ அதற்கு மேற்பட்ட பனுவல்களில் நடைமுறை படுத்தப்படும் (கொள்கை அடிப்படையில் சொற்களின் பைகள், அதாவது இயல்பாக நேரும் சொற்கள் கையாளப்படாலாம்). நிறுத்தற்குறிகளை விட்டுவிட்டால், நாம் பனுவல் T என்பதை சொற்களின் வரிசையாகப் $\{w_1, w_2,\ldots,w_n\}$ பார்க்கலாம்; மேலும் நாம் முறையாக சொற்பொருண்மை மயக்கநீக்கத்தை T என்பதில் உள்ள எல்லா அல்லது சில சொற்களின் பொருத்தமான பொருண்மையை/பொருண்மைகளைத் ஒதுக்கித்தரும் செயல்பாடாக விளக்க இயலும்; அதாவது சொற்களிலிருந்து பொருண்மைகளுக்கு A பொருத்ததைக் கண்டுபிடிக்க இயலும்; இதன் படி $A(i) \subseteq SensesD(w_i)$, இங்கு SensesD(wi) என்பது wi என்ற சொல்லுக்கு ஒரு அதிகராதி Dஇல் குறியாக்கம் செய்யப்பட்ட பொருண்மைகளின் குழுமம் மற்றும் A(i) என்பது சூழல் Tஇல் பொருத்தமான wiஇன் துணைக் குழுமம் ஆகும். இந்தப் பொருத்தம் A ஒவ்வொரு சொல்லுக்கும் ஒன்றுக்கும் மேற்பட்ட அர்த்தங்களை ஒதுக்கும் $wi \in T$; எடுத்துக்காட்டானதாக மிகப் பொருத்தமான பொருண்மை மட்டும் தேர்ந்தெடுக்கப்படும், அதாவது $|(A(i)| = 1$.

சொற்பொருண்மை மயக்கநீக்கம் ஒரு பாகுபடுத்தும் செயல்பாடாகப் (classification task) பார்க்கப்படும்: சொற்பொருண்மைகள் வகுப்புகளாகும் (classes); மற்றும் சூழலிலிருந்தும் வெளி அறிவு மூலங்களிலிருந்தும் கிடைக்கிற சான்று அடிப்படையில் ஒன்றோ அதற்கு கூடுதலோ வகுப்புகளுக்கு ஒரு சொல்லின் நேர்வை ஒதுக்கத் தானியக்க வகைப்படுத்தும் நெறிமுறை பயன்படுத்தப்படுகின்றது. பிற பாகுபடுத்தும் செயற்பாடுகள் இயற்கை மொழி ஆய்வுக் களத்தில் ஆயப்படுகின்றது: சொல்வகைப்பாடு அடையாளப்படுத்தல் (அதாவது சூழலில் இலக்குச் சொற்களுக்கு சொல்வகைப்பாடு ஒதுக்குகை), பெயரிடப்பட்ட இருப்புப்பொருள் தீர்மானம் (named entity resolution) (அதாவது இலக்கு பனுவல்களுக்கு முன்வரையறை விளக்கம் செய்யப்பட்ட புலக்குறிப்புகளின் ஒதுக்குகை) போன்றன. இந்தச் செயல்பாடுகளுக்கும்

சொற்பொருண்மை மயக்கநீக்கத்திற்கும் இடையிலான ஒரு முக்கியமான வேறுபாடு முந்தையது ஒரு தனியான முன்வரையறைவிளக்கம் செய்யப்பட்ட வகுப்புகளின் குழுமம் (சொல்வகைப்பாடு/parts of speech, வகைப்பாடுகள்/categories, போன்றன) ஆகும்; ஆனால் பிந்தையதில் வகுப்புகளின் குழுமம் வகைப்படுத்தப்பட வேண்டிய சொல் அடிப்படையில் மாறுகின்றது. இதன் அடிப்படையில் சொற்பொருண்மை மயக்கநீக்கம் n தனிப்பட்ட வகைப்பாட்டுச் செயற்பாடுகளைக் கொண்டதாகும்; இதில் n என்பது சொற்களஞ்சியத்தின் (lexicon) அளவாகும்.

இனப்பொதுமையான சொற்பொருண்மை மயக்கநீக்கத்தின் செயல்பாட்டின் இரண்டு மாறிகளை வேறுபடுத்த இயலும்:

சொல்சார் மாதிரி (அல்லது இலக்குச் சொற்பொருண்மை மயக்கநீக்கம்) (Lexical sample (or targeted WSD)): இதில் ஒரு வாக்கியத்திற்கு ஒன்று எனப் பொதுவாக நேர்கின்ற இலக்குச் சொற்களின் குழுமத்தைப் பொருண்மை மயக்கநீக்கம் செய்ய வேண்டி ஒரு ஒழுங்குமுறை தேவைப்படும். கண்காணிக்கப்பட்ட ஒழுங்குமுறைகள் இந்தப் பின்னணியமைப்பில் எடுத்துக்காட்டாகச் செயல்படுத்தப்பட்டுள்ளது; எனென்றால் அவற்றைக் கை-புலக்குறிப்பு செய்யப்பட்ட எடுத்துக்காட்டுகளைப் (பயிற்சிக் குழுமம் (training set)) பயன்படுத்தி பயிற்சி தரவியலும்; அதன்பின்னர் புலக்குறிப்பு செய்யப்படாத எடுத்துகாட்டுகளை (பரிசோதனைக் குழுமம் (test set)) வகைப்படுத்தப் பயன்படுத்தலாம்.

எல்லா-சொற்கள் சொற்பொருண்மை மயக்கநீக்கம் (All-words WSD): இதில் ஒழுங்குமுறைகள் ஒரு பனுவலில் உள்ள எல்லாத் திறந்த வகுப்புச் சொற்களையும் (அதாவது பெயர்கள், வினைகள், பெயரடைகள், வினையடைகள்) பொருண்மைமயக்கநீக்கம் செய்ய எதிர்பார்க்கப்படுகின்றது. இந்தச் செயல்பாடு விரிந்த-உள்ளடக்கப்பரப்பு ஒழுங்குமுறைகளை (wide-coverage system) வேண்டும். இதன்விளைவாகச் சுத்தமான கண்காணிக்கப்பட்ட ஒழுங்குமுறைகள் தரவுக் குறைவின் சிக்கலால் செயற்றிறமாகப் பாதிக்கப்படலாம்; ஏனென்றால் ஆர்வமூட்டும் மொழியின் முழுச் சொற்கஞ்சியத்தையும் உள்ளடக்கும் போதுமான அளவு பயிற்சிக் குழுமம் கிடைப்பது சாத்தியமல்ல. மாறாக அறிவு-சாய்வு ஒழுங்குமுறைகள் (knowledge-lean systems) போன்ற பிற அணுகுமுறைகள் முழு உள்ளடக்கப்பரப்பு மூலவளங்களைச் சார்ந்திருக்கின்றன; அவைகள் கிடைப்பது உறுதிசெய்யப்படவேண்டும்.

நாம் இப்பொழுது சொற்பொருண்மை மயக்கநீக்கத்தின் நான்கு தனிமங்களைப் பார்ப்போம்: சொற்பொருண்மையின் தேர்வு (selection of word sense) (அதாவது வகுப்புகள்), வெளி அறிவு மூலங்களின் பயன்பாடு (the use of external knowledge sources), சூழலின் உருப்படுத்தம் (the representation of context) மற்றும் ஒரு தானியியக்க வகைப்படுத்தும் நெறிமுறையின் தேர்வு (the selection of an automatic classification method).

4.1.1 சொற்பொருண்மைகளின் தேர்வு

ஒரு சொற்பொருண்மை ஒரு சொல்லின் பொதுப்படையாக ஏற்றுக்கொள்ளப்பட்ட பொருண்மையாகும். எடுத்துக்காட்டாக, பின்வரும் இரு வாக்கியங்களைக் கருத்தில் கொள்ளவும்:

அ. She chopped the vegetable with a chef's *knife*.

ஆ. A man was beaten and cut with a *knife*.

knife என்ற சொல் மேற்கண்ட வாக்கியங்களில் இரு வேறுபட்ட பொருண்மைகளில் பயன்படுத்தப்படுகின்றது: ஒன்று கருவி (a tool) மற்றொன்று ஓர் ஆயுதம் (a weapon). இந்த இரு பொருண்மைகளும் அவை ஒரே பொருளை (object) குறிப்பிடுவது சாத்தியமானதால் அவைத் தெளிவாகத் தொடர்புபடுத்தப்பட்டுள்ளன; இருப்பினும் பொருளின் விருப்பப் பயன்பாடுகள் வேறானவை. இந்த எடுத்துக்காட்டுகள் ஒரு சொல்லின் பொருண்மைத் தெரிவடையை (sense inventory) நிர்ணயிப்பது தான் சொற்பொருண்மை மயக்கநீக்கத்தில் ஒரு முக்கிய சிக்கல் எனத் தெளிவாக்குகின்றது: நாம் வாக்கியம் அ மற்றும் ஆ என்பவைகளில் உள்ள *knife* என்பதன் இரு நேர்வுகளுக்கு வேறுபட்ட வகுப்புகளை ஒதுக்க விரும்புகிறோமா?

ஒரு பொருண்மைத் தெரிவடை (sense invertory) ஒரு சொல்லின் பொருண்மையின் பரப்பெல்லையை அதன் பொருண்மைகளாகப் பிரிக்கின்றது. சொற்களின் பொருண்மைகளை எளிதில் வெவ்வேறாகப் பிரிக்க இயலாடாது; அதாவது ஒவ்வொன்றும் தனிப்பட்ட பொருண்மையைக் குறியாக்கம் செய்யும் பதிவுகளின் முற்றுப்பெற்ற வெவ்வேறான குழுவாகச் சுருக்கவியலும். இந்தக் கடினத்தின் முக்கியமான காரணம் மொழி மாற்றத்திற்கும் பொருள்கோளுக்கும் மரபுரிமையாக ஆட்படுத்தப்பட்டுள்ளது என்ற உண்மையிலிருந்து தோன்றுகின்றது. மேலும், தரப்பட்ட சொல்லுக்கு ஒரு பொருண்மை இறுதியுறும்போது அடுத்த பொருண்மை தொடங்குகின்றது என்று வாதிடலாம். எடுத்துக்காட்டாக *knife* என்ற சொல்லுக்கு

படம் 1-இல் தரப்பட்டுள்ள/கூறப்பட்டுள்ள பொருண்மைத் தெரிவடைவைக் கருத்தில் கொள்ளவும்.

படம்1: Knife/கத்தி என்ற பெயர்ச்சொல்லுக்குப் பட்டியலிடும் பதிவின் எடுத்துக்காட்டு

Knife n.1. a cutting tool composed of a blade with a sharp point and a handle. 2. an instrument with a handle and blade with a sharp point used as a weapon.

படம்2: knife/கத்தி என்ற பெயர்ச்சொல்லுக்கு ஆக்கமுறைப் பதிவின் எடுத்துக்காட்டு

Knife

TYPESTR = [ARG1 = [x] artifact tool]

ARGSTRU
- D-ARG1 = [y] physical object
- D-ARG2 = [w] human
- D-ARG3 = [z] human
- D-E1 = [e1] transition
- D-E2 = [e2] process

QUALIA =
- FORMAL = [x]
- CONSTITUTIVE = {blade, handle,…
- TELIC = cut act ([e2], [w]. [x], [y])
- AGENTIVE = make act ([e1]. [z]. [x])

கத்தி

வகையமைப்பு = [பங்கெடுப்பாளர்1 = [x] செய்யப்பட்ட கருவி]

பங்கெடுப்பாளர் அமைப்பு =
- D- பங்கெடுப்பாளர் 1 = [y] பௌதிகப் பொருள்
- D- பங்கெடுப்பாளர் 2 = [w] மனிதவினம்
- D- பங்கெடுப்பாளர் 3 = [z] மனிதவினம்
- D-E1 = [e1] மாற்றம்
- D-E2 = [e2] செயற்பாங்கு

குணவமைப்பு=
- முறை = [x]
- உறுப்பு = {blade, handle,…
- செயல் = cut act ([e2], [w]. [x], [y])
- செயலி = make act ([e1]. [z]. [x])

நாம் “ஒரு இயந்திரத்தின் பாகமாக அமையும் ஒரு வெட்டுவாய் (a cutting blade forming part of a machine)” என்பதற்குக் கூடுதல் அர்த்தத்தைத் தெரிவடையில் சேர்க்கவேண்டுமா அல்லது முதல் பொருண்மை இந்தப் பொருண்மையைக் கொண்டிருக்கின்றதா? இந்த தெளிவின்மையின் விளைவால் வெவ்வேறான அகராதிகளில் வேறுபட்ட விருப்பத்தேர்வு செய்யப்படும்.

மேலும் அர்த்த வேறுபாட்டுகளின் தேவையான நுண்மை பயன்பாட்டைப் பொறுத்து அமையும். எடுத்துக்காட்டாக, இயந்திர மொழிபெயர்ப்பில் மொழியைக் கடந்து சொற் பொருண்மை மயக்கம் தக்கவைக்கப்படும்/பாதுகாக்கப்படும் (எ.கா. interest என்ற சொல்லின் பொருண்மை மயக்கம் ஆங்கிலம், இத்தாலியன், பிரஞ்சு என்ற மொழிகளில் தக்கவைக்கப்படும்). இதன் விளைவாக, இப்பொருண்மைகளை எண்ணிக்கணக்கிடுவது/பட்டியலிடுவது மேலோட்டமானதாகும்.

மனிதன் மொழியைப் பிரிந்துகொள்வதைப் பொருண்மை மயக்கம் பாதிக்காது என்றாலும், சொற்பொருண்மை மயக்கநீக்கம் கணினியியல் முறையில் சூழலில் சொல்லில் உள்ளுறையும் பொருண்மையை வெளிப்படுத்துவதை இலக்காகக்கொள்ளும். சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் புறவய மதிப்பீட்டையும் (objective evaluation) ஒப்பீடையும் சாத்தியமாக்க வேண்டி, பொருண்மைகள் பொருண்மைத் தொடரடையில் கணக்கிடப்படவேண்டும் (கணக்கிடும்/பட்டியலிடும் அணுகுமுறை; பார்க்க படம் 1). எல்லா மரபுரிமைசார்ந்த நூல் வடிவிலான மற்றும் இயந்திரம் படிக்கவியலும் அகராதிகள் பட்டியலிடும்/கணக்கிடும் அணுகுமுறையை (enumerative approach) மேற்கொள்ளுகின்றன.

இருந்தபோதிலும், பொருண்மை வேறுபாடுகளைத் தூண்டுவது பல கேள்விகளை எழுப்பும்: (எ.கா. பனுவல்களின் சேகரிப்பில் சான்றுகளின் மேல்) நன்றாக நுணுக்கப்பட்ட (fine-grained) அல்லது அரைகுறையாக நுணுக்கப்பட்ட பொருண்மைகளைத் தருவதா வேண்டாமா? (splitting vs. lumping பொருண்மை வேறுபாடுகள்), அகராதியில் பொருண்மைகளை ஒழுங்கமைப்பது எவ்வாறு? போன்றன. இந்த வாதங்களுக்கு விடையாக ஆக்கமுறை அணுகுமுறை (generative approach) என்ற புதிய அணுகுமுறை முன்மொழியப்பட்டது (Pustejovsky 1991); இதில் தொடர்புள்ள பொருண்மைகள் பொருண்மை உருவாக்கத்தில் ஒழுங்குகளைப் பிரதிபலிக்கும் விதிகளிலிருந்து உருவாக்கப்பட்டன. பிந்தைய அணுகுமுறைக்குத்

தரப்படுகின்ற கூடுதல் நியாயம், முன்நிறுவப்பட பொருண்மைகளின் குழுமத்திற்குள் ஒரு சொல்லின் எப்பொழுதும் மாறுகின்ற வெளிப்படுத்தும் தன்மையைக் கட்டுப்படுத்துவது சாத்தியமல்ல (Kilgarriff 1997, 2006). ஆக்கமுறை அணுகுமுறையில், பொருண்மைகள் குணப் பங்களிப்புகள் (qualia roles) அடிப்படையில் வெளிப்படுத்தப்படுகின்றன; அதாவது, ஒரு இருப்புப்பொருளைப் பற்றிய அடிப்படை அறிவை அமைப்பாக்கம் செய்யும் பொருண்மைக் கூறுகளால் (semantic features) வெளிப்படுத்தப்படுகின்றன. இப்பண்புக்கூறுகள் சொல் அலகுகளின் பொருண்மையை விளக்குவதற்கு அரிஸ்டாட்டிலின் அடிப்படைத் தனிமங்களிலிருந்து (Aristotle's Basic elements) தோன்றுகின்றது. படம் 2 knife என்ற பெயருக்கு ஒரு ஆக்கமுறை பதிவின் எடுத்துக்காட்டைக் காட்டுகின்றது. [(இந்த எடுத்துக்காட்டுகள் ஜான்ஸ்டன் மற்றும் புஸா (Jonsoton and Busa, 1996) என்போரால் தரப்பட்டதாகும்); புஸ்தெஜொவ்ஸ்கியையும் (Pustejovsky 1995) பார்க்கவும்.] நான்கு குணப் பங்களிப்புகள் தரப்பட்டுள்ளன: முறைப் பங்களிப்பு (formal role) (கத்தி என்பதன் உள்ளடக்குச் சொல்/ a superordinate of knife), உறுப்புப் பங்களிப்பு (constitutive role) (கத்தியின் பாகங்கள்/parts of knife), செயல் பங்களிப்பு (telic role) (கத்தியின் பயன்பாடு/ the purpose of a knife), செயலி பங்களிப்பு (agentive role) (கத்தியைப் பயன்படுத்துபவர்/who uses a knife). பங்களிப்புகளின் ஒருங்கிணைப்புகளின் உடனடிநிகழ்வு ஒரு பொருண்மையின் உருவாக்கத்தை அனுமதிக்கின்றது. ஆக்கப்பொருண்மையியல் அணுகுமுறையைப் பின்பற்றி புயுடெலார் (Buitellar 1998) CoreLex என்ற ஒரு மூலவளத்தின் உருவாக்கத்தை முன்மொழிந்தார்; இது எல்லா ஒழுங்கான தொடர்புடைய பொருண்மைகளைக் கண்டுபிடிக்கும் மற்றும் குறைவாகச் சிறப்பிக்கப்பட்ட பொருண்மை அடையாளப்படுத்தலை அனுமதிக்கும். தெளிவில்லாத பொருண்மை வேறுபாடுகளை (fuzzier sense distinctions) இலக்குசெய்யும் பிற அணுகுமுறைகள் பொருண்மை அறிமுகத்தின் நெறிமுறைகளை உட்படுத்தும்; மேலும் மொழியியல் பின்னணியில் மொழியியல் அளவீடுகள் (linguistic criteria) (Cruse 1986) அடிப்படையில் பொருண்மை மயக்கப் பரிசோதனைகளையும் உட்படுத்தும்.

ஆய்வுச் சமூகத்தில் கணக்கிடும்/பட்டியலிடும் அணுகுமுறையின் (enumerative approach) பரந்துபட்ட ஏற்றுக்கொள்கை காரணமாக நாம் பின்வருவதில் கணக்கிடும்/பட்டியலிடும் அணுகுமுறையை ஏற்றுகொள்வோம். இருப்பினும் சொற்பொருண்மையின் தெளிவில்லாத

கருத்துச்சாயல் (fuzzier notion) அடிப்படையிலான செயல்கள் இவ்வியல் முழுவதும் கூறப்படும். நாம் D என்ற செயல்பாடாக ஒரு அகராதியில் குறியாக்கம் செய்யப்பட்ட சொற்களுடன் தனியான பொருண்மை வேறுபாடுகளின் (discrete sense distinction) சேர்க்கையை முறையாக்கம் செய்கின்றோம்:

SensesD: L x POS $\rightarrow 2^C$,

இதில் L என்பது சொற்களஞ்சியம், அதாவது, அகராதியில் குறியாக்கம் செய்யப்பட்ட சொற்களின் குழுமம், POS = {n, a, v, r} என்பது திறந்த-வகுப்பு சொல்வகைப்பாடுகள் (open-class parts of speech) (முறையே பெயர்கள், பெயரடைகள், வினைகள், வினையடைகள்), C என்பது அகராதி D-இல் உள்ள கருத்துருப் புலக்குறிப்புகளின் முழுக் குழுமம் (2^C இதன் கருத்துருக்களின் பெருக்குத்தொகைக் குழுமத்தைக் குறிப்பிடும்).

இவ்வியல் முழுவதும் நாம் w எனற ஒரு சொல்லை w_p எனக் குறிப்பிடுவோம்; இதில் p என்பது அதன் சொல்வகைப்பாடு ($p \in POS$), அதாவது நமக்கு $w_p \in L \times POS$ இருக்கிறது. இவ்வாறு w_p என்ற சொல்வகைபாட்டுக்கு அடையாளப்படுத்தப்பட்ட சொல் தரப்படுகையில், நாம *SensesD*(*w*, *p*) என்பதை *SensesD*(*wp*) ஆகச் சுருக்குகின்றோம்; இது w_p என்பதன் பொருண்மைகளை அது சேர்ந்துவரும் சூழலைப் பொறுத்து w_p குறிப்பிடுவதாகக் கருதப்படும் வேறுபட்ட பொருண்மைகளின் குழுமத்தைக் குறியாக்கம் செய்யும். ஒரு சொல் சொல்வகைப்பாட்டுக்கு (POS) அடையாளப்படுத்தப்பட்டது என்ற ஊகம் நியாயமானது என்று நாம் கண்டுகொள்கிறோம்; ஏனென்றால் தற்காலச் சொல்வகைப்பாட்டு அடையாளப்படுத்திகள் (POS taggers) மிக உயர்ந்த நிறைவுடன் இவ்வகையிலான பொருண்மை மயக்கத்தைத் தீர்க்கும். நாம் w_p என்ற சொல் ஒரே ஒரு பொருண்மையை மட்டும் வெளிப்படுத்தினால் அதை ஒருபொருண்மையானது (monosemous) என்று கூறுகின்றோம்; அதாவது $|\ SensesD(w_p)\ | = 1$. எடுத்துக்காட்டாக, *well-being*$_n$ என்பது ஒரு ஒருபொருண்மை/monosemous சொல்லாகும்; ஏனென்றால் அது 'that of being comfortable, happy or healthy' என்ற ஒரு தனிப் பொருண்மையைக் குறிப்பிடுகின்றது. மறுதலையாக w_p என்பது கூடுதல் பொருண்மைகளை வெளிப்படுத்தினால் அது பல்பொருண்மையானது/polysemous (எ.கா. *race*$_n$ as a competition, as a contest of speed, as a taxonomic group, etc.). தொடர்பில்லாத பொருண்மைகளை வெளிப்படுத்தும் w_p ஏன்ற சொல்லின் பொருண்மைகள் ஒருசொல்போலியாகும்/homonymous

(எ.கா. *race*$_n$ as a contest vs. *race*$_n$ as a taxonomic group). இறுதியாக, w என்ற சொல்லின் p என்ற சொல்வகைப்பாடு கொண்ட *i* ஆவது சொற்பொருண்மையை w^i_p எனக் குறிப்பிடுகின்றோம் (பிற குறியீடுகளும் பயன்பாட்டில் உள்ளன, எ.கா. *w#p#i*.

4.1.2 புற அறிவு மூலங்கள்

சொற்பொருண்மை மயக்கநீக்கத்தின் அடிப்படைக் கூறு/பகுதி அறிவு ஆகும். அறிவு மூலங்கள் சொற்களுடன் பொருண்மைகளைத் தொடர்புபடுத்தத் தேவையான தரவுகளைத் தருகின்றன. அவை புலக்குறிப்பு செய்யப்படாத (unlabeled) அல்லது அடையாளப்படுத்தப்பட்ட (annotated) பனுவல்களின் தரவுத்தொகுதிகளிலிருந்து (corpora of texts) இயந்திரத்தால் படிக்கவியலும் அகராதிகள் (machine readable dictionaries (MRDs)), பொருட்புல அகராதிகள் (thesauri), பொருள்விளக்கச் சொற்கோவை (glossaries), மூலப்பொருண்மையியல் ஆய்வுகள்/மெய்ப்பொருள் மூலாய்வு (ontologies) மற்றும் பிற என வேறுபடலாம்.

அமைப்பாக்கம் செய்யப்பட்ட மூலவளங்கள் (structured resources):

பொருட்புல அகராதிகள்: இவை சொற்களுக்கு இடையேயுள்ள உறவுகள் பற்றிய தகவல்களைத் தருகின்றன: ஒருபொருள்பன்மொழியம் (synonymy) (எ.கா. car$_n$ என்பது motorcar$_n$ என்பதன் ஒருபொருள்பன்மொழியாகும்), எதிர்மொழியம் (antonymy) (எதிரிடையான பொருண்மைகளை உருப்படுத்தம் செய்யும், எ.கா. *ugly*$_a$ என்பது *beautiful*$_a$ என்பதன் எதிர்மொழியாகும்), மற்றும் சாத்தியமான பிற உறவுகள் [Kilgarriff and Yallop 2000]. சொற்பொருண்மை மயக்கநீக்கக் களத்தில் மிகப் பரவலாகப் பயன்படுத்தப்படும் பொருட்புல அகராதி Roget's International Thesaurus (Roget 1911) என்பதாகும். இப்பொருட்புல அகராதியின் மிக அண்மைக்காலப் பதிப்பு 2,50,000 சொற்பதிவுகள் ஆறு வகுப்புகளிலும், 1000 வகைப்பாடுகளிலும் ஒழுங்குபடுத்தப்பட்டுள்ளன. சில ஆய்வாளர்கள் *Macquarie Thesaurus* [Bernard 1986] என்பதைப் பயன்படுத்துகின்றனர்; இது 2,00,000-க்கு மேற்பட்ட ஒருபொருள் பன்மொழிகளை குறியாக்கம் செய்துள்ளது.

இயந்திரத்தால் படிக்கவியலும் அகராதிகள்: இவை 1980களில் அகராதிகள் மின்வடிவில் கிடைக்கும்படி செய்யப்பட்ட போதே இயற்கைமொழி ஆய்விற்குப் பிரபலமான அறிவு மூலமாக மாறியுள்ளன; இவைகளில் பின்வருவன முக்கியமானவைகளாகும்: Collins English Dictionary, the Oxford Advanced Learner's Dictionary of Current English, the Oxford Dictionary of English

[Soanes and Stevenson 2003], and the Longman Dictionary of Contemporary English (LDOCE) [Proctor 1978]. பிந்தையன இயற்கைமொழி ஆய்வுச் சமூகத்தால் மிகப் பரவலாகப் பயன்படுத்தப்பட்ட இயந்திரத்தால் படிக்கவியலும் அகராதிகள் ஆகும் (பார்க்க Wilks et al. [1996]. சொல்வலை [WordNet [Miller et al. 1990; Fellbaum 1998] தற்போது ஆங்கிலத்தில் சொற்பொருண்மை மயக்கநீக்கத்திற்கு கூடுதலாகப் பயன்படுத்தப்படும் மூலவளமாகும். சொல்வலைகள் சாதாரண இயந்திரத்தால் படிக்கவியலும் அகராதிகளைவிட ஒரு படி முன்னால் இருப்தாகப் பெரும்பாலும் கருதப்படுகின்றது; இதன் காரணமாகச் சொல்வலை ஒரு கணினிவியல் சார் சொற்களஞ்சியம் எனச் சாதாரணமாக வரையறைவிளக்கம் செய்யப்படுகின்றது.

மூலப்பொருண்மையியல் ஆய்வுகள் (Ontologies): இவை ஆர்வமுள்ள குறிப்பிட்ட பொருட்புலங்களின் (specific doimains of interest) கருத்துருவாக்கங்களின் சிறப்பீடுகளாகும் (Gruber 1993); இவை பெரும்பாலும் ஒரு வகைப்பாட்டியலையும் (taxonomy) ஒரு குழும பொருண்மை உறவுகளையும் உள்ளடக்கி இருக்கும். இந்நிலையில் சொல்வலையையும் அதன் நீட்சிகளையும் மற்றும் சொல்வலையை மீண்டும் ஒழுங்படுத்தும் மற்றும் கருத்துருவாக்கம் செய்யும் முயற்சியான ஒமேகா மூலப்பொருண்மையியல் ஆய்வு [Philpot et al. 2005], சுமோ உயர்நிலை மூலப்பொருண்மையியல் ஆய்வு (the SUMO upper ontology) [Pease et al. 2002] போன்றவற்றை மூலப்பொருண்மையியல் ஆய்வுகளாகக் கருதவியலும். பொருட்புல நோக்கில்/அடிப்படையில் (domain oriented) ஒரு முயற்சி ஒருங்கிணைக்கப்பட்ட மருத்துவ மொழி ஒழுங்கமைப்பாகும் (Unified Medical Language System (UMLS)) [McCray and Nelson 1995]; இது மருத்துவக் கருத்துருக்களின் வகைப்பாட்டாக்கத்தைத் தரும் ஒரு பொருண்மை வலையமைப்பை (semantic network) உட்படுத்தும்.

அமைப்பாக்கம் செய்யப்படாத மூலவளங்கள் (unstructured resources):

தரவுத்தொகுதிகள் (corpora): இவை கற்கும் மொழி மாதிரிகளாகப் (learning language models) பயன்படுத்தப்படும் பனுவல்களின் சேகரிப்புகளாகும். தரவுத்தொகுதிகள் பொருண்மை-அடையாளப்படுத்தப்பட்டவைகளாகவோ (sense-annotated) கச்சாவாகவோ (பொருண்மை-அடையாளப்படுத்தப் படாதவைகளாகவோ (raw/unlabelled)) இருக்கவியலும். இவ்விரு மூலவளங்களும் சொற்பொருண்மை மயக்கநீக்கத்தில் முறையே கண்காணிக்கப்பட்ட

அணுகுமுறையிலும் (supervised approach) கண்காணிக்கப்படாத அணுகுமுறையிலும் (unsupervised approach) பயன்படுத்தப்பட்டுகின்றன.

கச்சாத் தரவுத்தொகுதிகள் (raw corpora): பிரவுன் தரவுத்தொகுதி (the Brown Corpus) [Kucera and Francis 1967] 1961-இல் ஐக்கிய அமேரிக்காவில் (Uinted States of America) வெளியிடப்பட்ட பனுவல்களின் சமநிலையான சேகரிப்பாகும்; பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி (the British National Corpus (BNC)) [Clear 1993] ஆங்கில மொழியின் எழுத்துவடிவ மற்றும் பேச்சுவடிவ மாதிரிகளின் 100 மில்லியன் சொற்களின் சேகரிப்பாகும் (இது பெரும்பான்மையும் சொல் நிகழ்வெண்களைச் சேகரிக்கவும் சொற்களுக்கு இடையிலான இலக்கண உறவுகளைக் கண்டுபிடிக்கவும் பயன்படுத்தப்படுகின்றது); வால்ஸ்ட்ரீட் ஜர்னல் தரவுத்தொகுதி/the Wall Street Journal (WSJ) corpus [Charniak et al. 2000], வால்ஸ்ட்ரீட் ஜர்னலிலிருந்து கிட்டத்தட்ட 30 மில்லியன் சொற்களின் சேகரிப்பு; அமேரிக்கன் தேசியத் தரவுத்தொகுதி/the American National Corpus [Ide and Suderman 2006] அமேரிக்கப் பேச்சு ஆங்கிலத்தின் 22 மில்லியன் சொற்களை உள்ளடக்கும்; the Gigaword Corpus செய்தித்தாள் பனுவலின் 2 பில்லியன் சொற்களின் சேகரிப்பு [Graff 2003] ஆகும்; மற்றும் பிற.

பொருண்மை அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் (sense-annotated corpus): SemCor [Miller et al. 1993] என்ற மிகப்பெரிய மற்றும் கூடுதல் பயன்படுத்தப்படுகின்ற பொருண்மை அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி; இது சுமார் 234,000 பொருண்மை அடையாளங்கள் உள்ள 352 பனுவல்கள் கொண்டது; மல்ட்டிசெம்கோர்/MultiSemCor [Pianta et al. 2002] என்ற ஒரு ஆங்கில-இத்தாலிய இணைத் தரவுத்தொகுதி (English-Italian parallel corpus); இது சொல்வலையின் ஆங்கிலம் மற்றும் இத்தாலியன் பதிப்புகளின் அர்த்தங்களால் அடையாளப்படுத்தப்பட்டுள்ளது; லைன்-ஹார்ட்-செர்வ் தரவுத்தொகுதி/line-hard-serve corpus [Leacock et al. 1993] என்ற 4000 பொருண்மை அடையாளப்படுத்தப்பட்ட (line, hard, serve) (முறையே பெயர், பெயரடை மற்றும் வினை) என்ற இம்மூன்று சொற்களின் எடுத்துக்காட்டுகளைக் கொண்டது; இன்ரஸ்ட் தரவுத்தொகுதி (*interest* corpus) [Bruce and Wiebe 1994] *interest* என்ற பெயரின் 2369 பொருண்மை-புலப்படுத்தப்பட்ட எடுத்துக்காட்டுகளைக் கொண்டது; டி.எஸ்.ஓ. தரவுத்தொகுதி (DSO corpus) [Ng and Lee 1996] என்பது சிங்கப்பூரின் தற்காப்பு அறிவியல் நிறுவனத்தால் உருவாக்கப்பட்டது; இது பிரவுண் மற்றும் வால் ஸ்ட்ரீட்

இதழ் தரவுத்தொகுதிகளிலிருந்து (Brown and Wall Street Journal corpora) 191 சொற்களின் 192,800 பொருண்மை அடையாளப்படுத்தப்பட்ட டோக்கன்களை உள்ளடக்கியது; the Open Mind Word Expert data set [Chklovski and Mihalcea 2002]: ஒரு ஒருங்கிணைந்த முயற்சியால் இணையதள பயன்பாட்டாளர்களால் பொருண்மையியல் அடிப்படையில் அடையாளப்படுத்தப்பட்ட 288 பெயர்களின் நேர்வுகள் கொண்ட வாக்கியங்களின் ஒரு தரவுத்தொகுதி; நான்கு மதிப்பீட்டு நடவடிக்கைகளிருந்தான பொருண்மையியல் அடிப்படையில் அடையாளப்படுத்தப்பட்ட தரவுதொகுதிகளான சென்ஸ்வல் மற்றும் செமெவல் தரவுக் குழுமங்கள் (Senseval and Semeval data sets). இந்த எல்லாத் தரவுத் தொகுதிகளும் சொல்வலை பொருண்மை தெரிவடையின் வேறுபட்ட மாதிரிகளால் அடையாளப்படுத்தப்பட்டுள்ளன; (LSOCE பொருண்மைகளால் அடையாளப்படுத்தப்பட்ட) இன்ட்ரஸ்ட் தரவுத்தொகுதியும் (interest corpus) HECTOR பொருண்மைத் தெரிவடையால் பொருண்மை-புலக்குறிப்பு செய்யப்பட்ட Oxford University Press/Digital project [Atkins 1993]-இன் கூட்டு முயற்சியின் விளைவான ஒரு சொற்களஞ்சியமும் தரவுத்தொகுதியுமான சென்ஸ்வல்-1 தரவுத்தொகுதியும் (Senseval-1 corpus) விதிவிலக்குகளாகும்.

சேர்ந்துவருகை மூலவளம் (collocational resources): இது சொற்கள் பிற சொற்களுடன் வழக்கமாகச் சேர்ந்துவரும் போக்கைப் பதிவு செய்கின்றது. எடுத்துக்காட்டுகள்: Word Sketch Engine, JustTheWord, The British National Corpus collocations, the Collins Cobuild Corpus Concordance, போன்றவை. சமீபகாலத்தில் பனுவல் சேர்ந்துவருகை நேர்வுகளின் மிகப்பெரிய தரவுக்குழுமம் (dataset) வெளியிடப்பட்டது; இது சொற்பொருண்மை மயக்கநீக்கச் சமூகத்தில் மிகவும் புகழ் பெற்றது; இது Web1T corpus [Brants and Franz 2006] என்பதாகும். இந்தத் தரவுத்தொகுதி இணைய வலையிலிருந்து (Web) ஆக்கப்பட்ட ஒரு முன்றுமில்லியன் சொல் தரவுத்தொகுதியில் ஐந்து சொற்கள்வரையுள்ள கோவைகளின் நிகழ்வெண்களைக் கொண்டதாகும்.

பிற மூலவளங்கள்: சொல் நிகழ்வெண் பட்டியல்கள் (word frequency lists), நிறுத்தப் பட்டியல்கள் (stoplists) (வேறுபடுத்தப்படாத a, an, the போன்ற சொற்கள்), பொட்புலப் பட்டியல்கள் (domain lists) [Magnini and Cavagli`a 2000] என்பன.

பின்வரும் துணைப் பகுதிகளில் இக்களதில் பரவலாகப் பயன்படுத்தப்படும் இரண்டு அறிவு மூலங்களைப் பற்றி விளக்கங்கள் தரப்படும்: சொல்வலை (wordNet) மற்றும் செம்கோர் (Semcor).

4.1.2.1 சொல்வலை

சொல்வலை (WordNet) [Miller et al. 1990; Fellbaum 1998] பிரின்ஸ்டன் பல்கலைக்கழகத்தில் (Prinsten University) உருவாக்கப்பட்டுப் பராமரிக்கப்படுகிற உளமொழியியல் கொள்கைகள் அடிப்படையிலான ஆங்கிலக் கணினியியல்சார் சொற்களஞ்சியம் (Computational Lexicon) ஆகும். இது ஒருபொருள்பன்மொழிகளின் (synsets) குழுமங்களின் அடிப்படையில் கருத்துருக்களைக் குறியனாக்கம் செய்கின்றது. இதன் அண்மைக்கால மாதிரியான சொல்வலை 3.0 155,000 சொற்களை 117,000 ஒருபொருள்பன்மொழிகளாக ஒழுங்குபடுத்தியுள்ளது. எடுத்துக்காட்டாக, automobile என்ற கருத்துரு பின்வரும் ஒருபொருள்பன்மொழியால் வெளிப்படுத்தப்பட்டுள்ளது. (உயர்நிலை எழுத்தும், கீழ்நிலை எழுத்தும் முறையே சொல்லின் பொருண்மை அடையாளமாகவும் சொல்வகைப்பாட்டு அடையாளமாகவும் இருக்கிறது.)

{car$^{1}_{n}$, auto $^{1}_{n}$, automobile $^{1}_{n}$, machine $^{4}_{n}$, motorcar $^{1}_{n}$}

நாம் ஒரு ஒருபொருள்பன்மொழியக் குழுமத்தை (தோராயமாக) ஒரே பொருண்மையை வெளிப்படுத்தும் எல்லாச் சொற்பொருண்மைகளின் ஒரு குழுமம் எனப் பார்க்கலாம். முன்னர் வந்த பகுதியில் அறிமுகப்படுத்தியுள்ள கருத்துச்சாயலின் படி, பின்வரும் செயல்பாடு ஒவ்வொரு சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட சொல் w_p-உடன் அதன் சொல்வலைப் பொருண்மைகளைச் சேர்க்கின்றது:

$Senses_{WN} : L \times POS \rightarrow 2^{SYNSETS}$,

இதில் SYNSETS என்பது சொல்வலையில் உள்ள முழு ஒருபொருள்பன்மொழியக் குழுமம். எடுத்துகாட்டு:

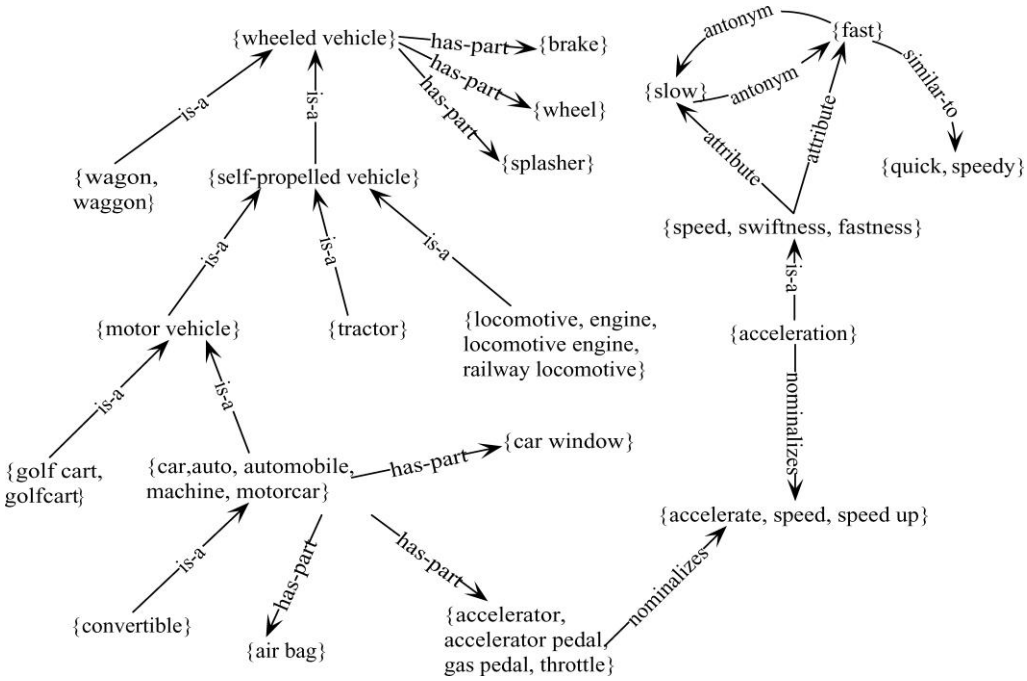
SensesWN(car$_n$) = {{car$^{1}_{n}$, auto$^{1}_{n}$, automobile$^{1}_{n}$, machine$^{4}_{n}$, motorcar$^{1}_{n}$},

{car $^{2}_{n}$, rail car $^{1}_{n}$, rail way car $^{1}_{n}$, rail road car $^{1}_{n}$} ,

{cable car $^{1}_{n}$, car $^{3}_{n}$},

{car $^{4}_{n}$, gondola $^{3}_{n}$},

{car $^{5}_{n}$, elevator car $^{1}_{n}$}}.



படம் 3: சொல்வலை பொருண்மை வலைப்பின்னலின் (semantic network) ஒரு பாகம்

நாம் ஒவ்வொரு சொற்பொருண்மையும் ஒரு தனி ஒருபொருள்பன்மொழியக் குழுமத்தை அடையாளம் காண்பதாகக் காண்கின்றோம். எடுத்துக்காட்டாக, *car*$^{1}{}_{n}$ தரப்படுகையில் அதற்குப் பொருத்தமான {{car$^{1}{}_{n}$, auto$^{1}{}_{n}$, automobile$^{1}{}_{n}$, machine$^{4}{}_{n}$, motorcar$^{1}{}_{n}$} என்ற ஒருபொருள்பன்மொழியக்குழுமம் ஒட்டுமொத்தமாகத் தீர்மானிக்கப்படுகின்றது. படம் 3-இல் நாம் car$^{1}{}_{n}$ என்பதைக் கொண்டிருக்கும் சொல்வலை பொருண்மை வலையமைப்பின் ஒரு பகுதியைக் கூறுகின்றோம். ஒவ்வொரு ஒருபொருள்பன்மொழிக்கும் சொல்வலைப் பின்வரும் தகவலைத் தருகின்றது:

பொருள்விளக்கம் (gloss): இது ஒரு குழுமப் பயன்பாட்டு எடுத்துக்காட்டுகளுடன் கூடிய ஒருபொருள்பன்மொழியக் குழுமத்தின் ஒரு பனுவல்சார் வரையறைவிளக்கம் ஆகும். (எ.கா. *car* $^{1}{}_{n}$ என்பதன் பொருள்விளக்கம் “a 4-wheeled motor vehicle; usually propelled by an internal combustion engine; ‘he needs a car to get to work’ ”).

சொல்லன்சார் மற்றும் பொருண்மைசார் உறவுகள் (Lexical and semantic relations): இவைகள் முறையே சொற்பொருண்மைகள் மற்றும் ஒருபொருள்பன்மொழிக் குழுமங்களின் இணைகளைத் தொடர்புபடுத்துகின்றது; பொருண்மைசார் உறவுகள் ஒருபொருள்பன்மொழியக் குழுமங்களுடன் முழுநிறைவாக (அதாவது ஒழுபொருள்பன்மொழியக் குழுமத்தின் உறுப்பினர்களால்)

அவற்றிற்குப் பயன்படுகின்றது; சொல்சார் உறவுகள் ஒருபொருள்பன்மொழியக் குழுமங்களுடன் முறையே உட்படுத்தப்பட்டுள்ள சொற்பொருண்மைகளை இணைக்கின்றது. பிந்தையதில் பின்வருவன உட்படும்:

எதிர்மொழியம் (antonymy): Y என்பது X-க்கு எதிரிடையான கருத்துருவை வெளிப்படுத்தினால் X என்பது Y என்பதன் எதிர்மொழி ஆகும் (எ.கா. good[1] $_{a}$ என்பது bad[1]$_{a}$ என்பதன் எதிர்மொழியாகும்). எதிர்மொழியம் எல்லாச் சொல்வகைப்பாட்டிற்கும் வரும்.

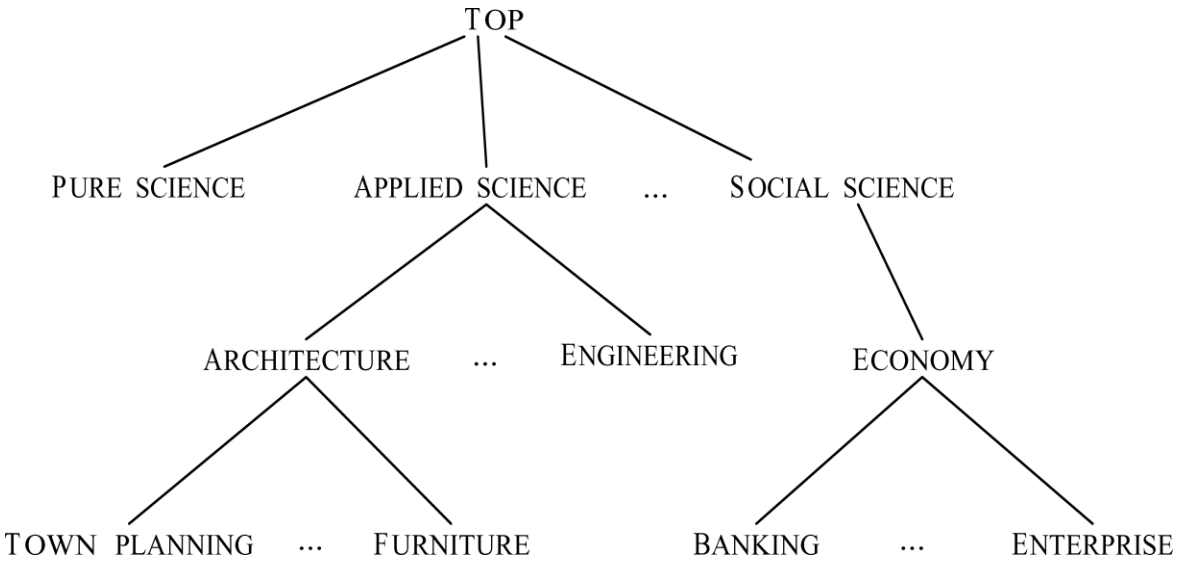
சார்பு மொழியம் (Pertainymy): X என்ற ஒரு பெயரடை Y என்ற ஒரு பெயருடன் (அரிதாக மற்றொரு பெயரடையுடன்) சார்புடையது (தொடர்புடையது) என்று வரையறை விளக்கம் செய்யலாம் (எ.கா. dental [1]$_{a}$ என்பது tooth [1]$_{n}$ என்பதுடன் தொடர்புடையது).

பெயராக்கம் (Nominalization): X என்ற ஒரு பெயர் Y என்ற ஒரு வினையைப் பெயராக்கம் செய்கின்றது (எ.கா. service[2]$_{n}$ என்பது serve[4]$_{v}$ என்பதைப் பெயராக்கம் செய்கின்றது).

பொருண்மைசார் உறவுகளில் பின்வருவன உள்ளன:

உள்ளடக்கு மொழியம் (hypernymy) ['ஒரு வகை' (kind-of) உறவு அல்லது 'ஆக இரு' உறவு (is-a)]: X என்பது Y என்பதன் ஒருவகையாக இருந்தால் Y என்பது X என்பதன் உள்ளடக்கு மொழியம் ஆகும். (எ.கா. motor vehicle[1]$_{n}$ என்பது car[1]$_{n}$ என்பதன் உள்ளடக்கு மொழியம் ஆகும்). உள்ளடக்கு மொழியம் பெயர்கள்சார் அல்லது வினைகள்சார் ஒருபொருள்பன்மொழியக் குழுமத்தின் இணைகளுக்கிடையில் வரும்.

படம் 4: சொல்வலை பொருட்புலப் புலக்குறிப்புகளின் வகைப்பாட்டியல்



உள்ளடங்கு மொழியம் மற்றும் உட்படு மொழியம் (Hyponymy and troponymy): இவை முறையே பெயர்சார் மற்றும் வினைசார் ஒருபொருள்பன்மொழிக் குழுமங்களுக்கு உள்ளடக்கு மொழியத்தின் மறுதலைகள் ஆகும்.

சினைமொழியம் (Meronymy or part-of): Y என்பது X என்பதன் சினையாக (பாகமாக) இருந்தால் Y என்பது X என்பதன் சினைமொழி ஆகும். (எ.கா. flesh$^{3}_{n}$ என்பது fruit$^{1}_{n}$ என்பதன் சினைமொழியாகும்).

முழுமொழியம் (Holonymy): X என்பது Y என்பதன் பாகமாக இருந்தால் Y என்பது X என்பதன் முழுமொழியம் ஆகும். இது சினைமொழியத்தின் மறுதலையாகும்.

உள்ளுறை மொழியம் (entailment): X-ஐச் செய்வது Y-ஐச் செய்வதை வெளிப்படுத்தினால் Y என்ற வினை X என்ற வினையால் உள்ளுறை செய்யப்படுகிறது எனக்கூறலாம். (எ.கா. sonre$^{1}_{v}$ என்பது snotr$^{1}_{v}$-ஐ உள்ளுறை செய்கின்றது.

ஒற்றுமை (similarity): X Y beautiful$^{1}_{a}$ என்ற பெயரடை attractive$^{1}_{a}$ என்ற பெயரடையுடன் ஒற்றுமை உடையது.

அடை (attribute): Y என்ற பெயரடை ஒரு மதிப்பை வெளிப்படுத்தும் அடை X என்ற பெயர் ஆகும் (எ.கா. hot$^{1}_{a}$ என்பது temperature$^{1}_{n}$ என்பதன் மதிப்பாகும்).

மேலும் பார்க்கவும் (see also): இது பெயரடைகளுக்கு இடையில் உள்ள உறவுத்தன்மையின் ஒரு உறவாகும் (எ.கா. beautiful[1]a என்பது attractive[1]a என்பதுடன் 'மேலும் பார்க்கவும்' என்ற உறவு வழி தொடர்புகொண்டுள்ளது).

மக்னினி மற்றும் கவாக்லி (Magnini and Cavagli`a 2000) சொல்வலை ஒருபொருள்பன்மொழியக் குழுமங்களுக்குப் பொருட்புலப் புலக்குறிப்புகளின் ஒரு தரவுக் குழுமத்தை உருவாக்கியுள்ளனர். சொல்வலை ஒருபொருள்பன்மொழியக்குழுமங்கள் டியுவி தசமப்பின்னப் பாகுபாட்டிலிருந்து (Dewey Decimal Classification) (எ.கா. உணவு, கட்டிடக்கலை, விளையாட்டு, போன்றன/ FOOD, ARCHITECTURE, SPORT, etc) சுமார் 200 அடையாளங்களின் முன்வரையறுக்கப்பட்ட குழுமத்திலிருந்து ஒன்றோ அதற்கு மேலோ பொருட்புல புலக்குறிப்புகளாலும் பொருட்புலத் தகவல் இல்லாதபோது ஒரு இனப்பொதுப் புலக்குறிப்பிலிருந்து (generic label) (FACTOTUM) பகுதி தானியக்கமாக அடையாளப்படுத்தப்பட்டுள்ளது. புலக்குறிப்புகள் ஒரு படிநிலை அமைப்பில் (எ.கா. உளவியல் ஆய்வு உளவியல் பொருட்புலத்தின் ஒரு வகை ஆகும்). படம் 4 பொருட்புலத்தின் ஒரு பகுதியைக் காட்டுகின்றது.

ஆய்வுச் சமூகத்திற்குள் அதன் விரிந்த பரவல் கரணமாக, சொல்வலையை ஆங்கிலச் சொற்பொருண்மை மயக்கநீக்கத்திற்கு ஒரு உண்மையான அளவுகோலாகக் கருதலாம். அதன் வெற்றியைப் பின்பற்றி பல மொழிகளுக்குச் சொல்வலைகள் உருவாக்கப்பட்டு மூல பிரின்ஸ்டன் சொல்வலையுடன் (Princeton WordNet) தொடர்புபடுத்தப்பட்டுள்ளன. இத்திசையை நோக்கிய முதல் முயற்சி யூரோ சொல்வலைத் திட்டத்தின் (Euro WordNet Project) (Vossen 1998) சூழலில் மேற்கொள்ளப்பட்டுள்ளது; இது தேசீய சொல்வலைகளுக்கிடையில் மொழியிடை ஒழுங்குபடுத்தலைத் (interlingual alignment) தருகின்றது. இந்நாட்களில் வேறுபட்ட மொழிகளுக்குச் சொல்வலைகளை உருவாக்கவும் மேம்படுத்தவும் பராமரிக்கவும் பல முயற்சிகள் எடுக்கப்பட்டு வருகின்றன; MultiWordNet [Pianta et al. 2002], BalkaNet [Tufis et al. 2004] போன்றன இதற்கு எடுத்துக்காட்டுகளாகும்.

உலகச் சொல்வலைக் கழகம்/GlobalWordNet Association உலகிலுள்ள எல்லா மொழிகளின் சொல்வலைகளைப் பங்கிட்டுக்கொள்ளவும் தொடர்புபடுத்தவும் நிறுவப்பட்டுள்ளது. இதிட்டங்கள் பிற மொழிகளில் சொற்பொருண்மை மயக்கநீக்கத்தைச் சாத்தியமாக்குவதுடன்

இயந்திர மொழிபெயர்ப்புக்குச் சொற்பொருண்மை நீக்கத்தின் பயன்பாட்டைச் சாத்தியமாக்கவியலும்.

4.1.2.2 செம்கோர்/SemCor

செம்கோர்/SemCor (Miller et al. 1993) பிரவுண் தரவுத்தொகுதியின் (Brown Corpus) (Kucera and Francis 1967) துணைகுழுமமாகும்; இதன் பொருளடக்கச் சொற்கள் (content words) சொல்வலை தெரிவடையிலிருந்து (wordNet inventory) சொல்வகைப்பாட்டு அடையாளங்கள் (part-of-speech tags), சொல்லன்கள் (lemmas), சொற்பொருண்மைகள் (word senses) என்பனவற்றால் கைகளால் அடையாளப்படுத்தப்பட்டுள்ளன. செம்கோர்/SemCor 352 பனுவல்களால் உருவாக்கப்பட்டுள்ளது: 186 பனுவல்களில் திறந்த-வகுப்புச் சொற்கள் (பெயர்கள், வினைகள், பெயரடைகள், வினையடைகள்) இத்தகவல்களால் அடையாளப்படுத்தப்பட்டுள்ளன; மீதமுள்ள 166 பனுவல்களில் வினைகள் மட்டும் சொற்பொருண்மைகளால் பொருண்மையியல் நிலையில் அடையாளப்படுத்தப்பட்டுள்ளன.

ஒட்டுமொத்தமாக, செம்கோர்/SemCor பொருண்மையியல் நிலையில் அடையாளப்படுத்தப்பட்ட சொற்களின் கிட்டத்தட்ட 234,000 மாதிரியால் (sample) ஆனதாகும்; இது கண்காணிக்கப்பட்ட பின்னணியமைப்புகளில் பொருண்மைப் பாகுபடுத்திகளைப் பயிற்சிசெய்ய மிகப்பெரிய பொருண்மை அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியைக் கொண்டிருக்கின்றது. தரவுத்தொகுதியில் ஒரு பனுவலின் பகுதி படம் 5-இல் தரப்பட்டுட்ள்ளது.

படம் 5: SemCor பொருண்மைநிலையில் அடையாளப்படுத்தப்பட்ட தரவுதொகுதியின் பாகம்:

> As of Sunday$^{1}_{n}$ night$^{1}_{n}$ there was$^{4}_{v}$ no word^{2}n of a resolution1*n* being offered2*v*there1*r* to rescind1*v*the action1*n*. Pelham pointed out1*v*that Georgia1*n* voters1*n* last1*r* November1*n* rejected2*v*aconstitutional1*a* amendment1*n* to allow2*v* legislators1*n* to vote1*n* on pay1*n* raises1*n* for future1*a* Legislature1*n* sessions2*n*.

எடுத்துக்காட்டாக word$_{n}$ முதல் வாக்கியத்தில் #2 என்ற பொருண்மையால் அடையாளப் படுத்தப்பட்டுள்ளது; இது சொல்வலையில் “brief statement” என்று விளக்கப்பட்டுள்ளது. (இதை எடுத்துக்காட்டாக “a unit of language that native speakers can identify” என்று வரையறைவிளக்கம் தரப்பட்ட பொருண்மை #1 என்பதுடன் ஒப்பிடுக). மூல SemCor சொல்வலை 1.5 அடிப்படையில் அடையாளப்படுத்தப்பட்டுள்ளது. இருப்பினும், பொருத்தங்கள்

அண்மைகாலத்துப் பதிவுமாதிரிகளுடன் (version) (எ.கா. 2.0, 2.1, போன்றவற்றுடன்) செய்யப்பட்டுளது.

SemCor அடிப்படையில், ஒரு இருமொழிய தரவுத்தொகுதி பெந்திவொக்லி மற்றும் பியந்தா (Bentivogli and Pianta 2005) என்பவர்களால் உருவாக்கப்பட்டது: மல்ட்டிசெம்கோர்/MultiSemCor என்பது ஒரு ஆங்கிலம்/இத்தாலியன் இணைத் தரவுத்தொகுதி (parallel corpus) ஆகும்; இது சொல் நிலையில் வரிசைப்படுத்தப்பட்டுள்ளது; இது ஒவ்வொரு சொல்லுக்கும் அதன் சொல்வகைப்பாடு, அதன் சொல்லன், சொல்வலையின் ஆங்கில மற்றும் இத்தாலியன் மாதிரிகளிலிருந்து (MultiWordNet (Pianta et al 2002) ஒரு பொருண்மை என்பனவற்றைத் தருகின்றது. இத்தரவுத்தொகுதி சொல் நிலையில் செம்கோர்/SemCor என்பதன் இத்தாலியன் மொழிபெயர்ப்பை வரிசைப்படுத்தி உருவாக்கப்பட்டுள்ளது. பின்னர் செம்கோர்/SemCor-இலிருந்து அசல் சொல் பொருண்மை அடையாளங்கள் வரிசைப்படுத்தப்பட்ட இத்தாலியச் சொற்களுக்கு மாற்றப்படுகின்றன.

4.1.3 சூழலின் உருப்படுத்தம் (representation of context)

பனுவல் தகவலின் ஒரு அமைப்பாக்கம் செய்யப்படாத மூலமாக இருப்பதால் இதை ஒரு தானியக்க நெறிமுறைக்குப் பொருத்தமான உள்ளிடாக மாற்ற இது பெருபாலும் அமைப்பாக்கம் செய்யப்பட்ட வடிவமைப்புக்கு மாற்றப்படுகின்றது. இதுவரை உள்ளீடு செய்யப்பட்ட பனுவலில் ஒரு முற்செயற்பாங்கு (pre-processing) பெரும்பாலும் நிறைவேற்றப்படுகின்றது; இது மாதிரியாகப் (தேவையானதாக அல்ல) பின்வரும் நடவடிகளை உட்படுத்தும்:

சொல்லலகுகளாகப் பிரித்தல் (tokenization): இது பனுவலைச் சொல்லலகுகளின் (பெரும்பாலும் சொற்களின்) குழுமங்களாகப் பிரிக்கும் இயல்பாக்க நடவடிக்கையாகும்.

சொல்வகைப்பாட்டு அடையாளப்படுத்தல் (part-of-speech tagging): இது ஒவ்வொரு சொல்லுக்கும் ஒரு இலக்கண வகைப்பாட்டை ஒதுக்கும் செயற்பாங்காகும். எ.கா. “the/DT bar/NN was/VBD crowded/JJ,” இதில் DT, NN, VBD, மற்றும் JJ என்பன முறையே determiners, nouns, verbs, மற்றும் adjectives, என்பனவற்றின் அடையாளங்கள் ஆகும்.);

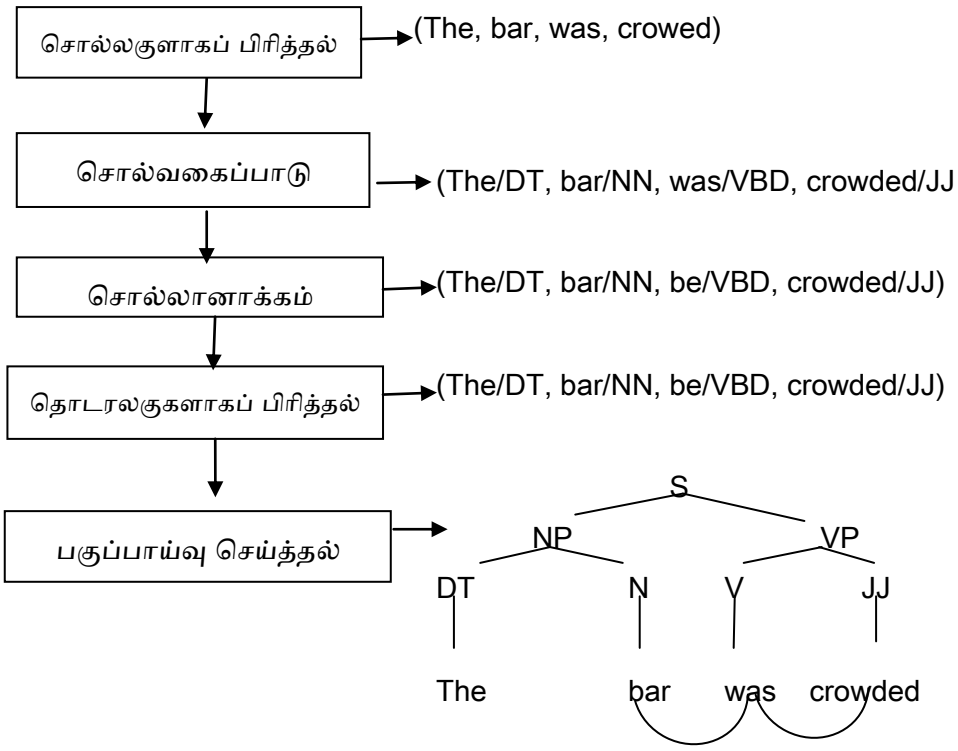
சொல்லனாக்கம் (lemmatization): இது மாற்றுருபன்களை (morphological variants) அவற்றின் அடிப்படை வடிவுக்குச் (base form) சுருக்குவதாகும்.

தொடரலகுகளாகப் பிரித்தல் (Chunking): இது ஒரு பனுவலைத் தொடரியல் அடிப்படையில் பொருத்தமான பாகங்களாகப் பிரிப்பதாகும். (எ.கா. [the bar]NP [was crowded]VP; இதில் NP-யும் VP-யும் முறையே noun phrase-யும் verb phrase-யும் குறிப்பிடும்).

பகுப்பாய்வுசெய்தல் (parsing): இது ஒரு வாக்கியத்தின் தொடரியல் அமைப்பை அடையாளம்கண்பதை நோக்கமாகக்கொண்டது [பொதுவாக வாக்கிய அமைப்பின் பகுத்துக்குறித்தக் கிளையமைப்பின் (parse tree) உருவாக்கத்தை உட்படுத்தும்.]

படம் 6: பனுவலின் முற்செயற்பாங்கு நடவடிக்கைகளின் எடுத்துக்காட்டு

The bar was crowded



படம் 6-இல் செயற்பாங்கின் ஒழுக்கின் (processing flow) எடுத்துக்காட்டு தரப்பட்டுள்ளது. பனுவலின் ஒரு பகுதியின் முற்செயற்பாங்கு நிலையின் விளைவாக (எ.கா. ஒரு வாக்கியம், ஒரு பத்தி, ஒரு முழு ஆவணம் போன்றன), ஒவ்வொரு சொல்லையும் வேறுபட்ட பண்புக்கூறுகளின் திசையன்/வெக்டார் (vector) அல்லது கூடுதல் அமைப்பாக்க வழிகளில், எடுத்துக்காட்டாக, சொற்களுக்கிடையே உள்ள உறவுகளின் ஒரு கிளையமைப்பாக (tree) அல்லது ஒரு வரைபடமாக (graph) உருப்படுத்தம் செய்யலாம். குறிப்புரை/மேற்கோள் தெரிவடைவிலிருந்து (reference

inventory) பொருத்தமான பொருண்மையைத் தேர்ந்தெடுக்கத் தானியக்க நெறிமுறைகளை அனுமதிக்கக் கூடுதல் அறிவு மூலவளங்களுடன் கூடிய சூழலில் ஒரு சொல்லின் உருப்படுத்தம் முக்கியமான ஆதரவாகும்.

சூழலை உருப்படுத்தம் செய்ய ஒரு குழுமப் பண்புக்கூறுகள் தெரிந்தெடுக்கப்படுகின்றன. இவை சொல்வகைப்பாட்டு அடையாளங்கள் (part-of-speech tags), இலக்கண உறவுகள் (grammatical relations), சொல்லன்கள் போன்ற முன்னர் கூறப்பட்ட முன்செயற்பாங்கு நடவடிக்கைகளிலிருந்து விளையும் தகவல்களை உட்படுத்தும். (ஆனால் இவற்றிற்கு மட்டும் எல்லைப்படுத்தப்பட்டுள்ளது என்று கூறவியலாது.) நாம் இப்பண்புக்கூறுகளைப் பின்வருமாறு குழுமலாம்.

வட்டாரப் பண்புகூறுகள் (local features): இவை ஒரு சொல் பயன்பாட்டின் வட்டாரச் சூழலை உருப்படுத்தம் செய்கின்றது; அதாவது சொல்வகைப்பாடு அடையாளங்கள் (part-of-speech tags), சொல் வடிவங்கள் (word forms), இலக்குச் சொல்லைப் பொறுத்து இருப்பிடங்கள் என்பனவற்றையும் உட்படுத்தி இலக்குச் சொல்லைச் சுற்றிய சிறு எண்ணிகையிலான சொற்களின் பண்புக்கூறுகளை உருப்படுத்தம் செய்யும்.

தலைப்புப் பண்புக்கூறுகள் (topical features): இது வட்டாரப் பண்புக்கூறுகளுக்கு முரணாக ஒரு பனுவலின் அல்லது கருத்தாடலின் பொதுத் தலைப்பை வரையறை விளக்கம் செய்கின்றது; இவ்வாறு இது பொதுவானசூழல்களை (எ.கா. சொற்களின் ஒரு சாளரம், ஒரு வாக்கியம், ஒரு தொடர், ஒரு பத்தி போன்றவை) உருப்படுத்தம் செய்கின்றது.

தொடரியல் பண்புக்கூறுகள் (syntactic features): இவை தொடரியல் குறிப்புகளையும் (syntactic cues) ஒரே வாக்கியத்திற்குள் இலக்குச் சொல்லுக்கும் பிற சொற்களுக்கும் இடையில் பங்கெடுப்பாளர்-தலைமை உறவுகளையும் (argument-head relations) உருப்படுத்தம் செய்யும். (இச்சொற்கள் வட்டார சூழலுக்கு வெளியேயும் வரலாம் என்பதைக் கவனத்தில் கொள்ளவும்.)

பொருண்மைசார் பண்புகூறுகள் (semantic features): இவை சூழல், பொருட்புலச், சுட்டிக்காட்டிகள் (domain indicators) போன்றவற்றில் வரும் சொற்களின் முன்னர் நிர்ணயிக்கப்பட்ட பொருண்மைகள் போன்ற பொருண்மையியல் பண்புக்கூறுகளை உருப்படுத்தம் செய்யும்.

இப்பண்புக்கூறுகளின் குழுமம் அடிப்படையில் ஒவ்வொரு சொல் நேர்வும் (பெரும்பாலும் ஒரு வாக்கியத்திற்குள்) ஒரு பண்புகூறு திசையனாக/வெக்டாராக (feature vector) மாற்றப்படவேண்டும். எடுத்துக்காட்டாக அட்டவணை 1 பின்வரும் வாக்கியங்களுக்கு ஒரு சாத்தியமான பண்புக்கூறு வெக்டாரின்/திசைநோக்கின் ஒரு எளிய எடுத்துக்காட்டை எடுத்துக்காட்டுகின்றது:

அட்டவணை 1: bank என்ற பெயரை இலக்காகக் கொண்ட இருவாக்கியங்களுக்குப் பண்புக்கூறு வெக்டர்களின்/திசைநோக்குகளின் எடுத்துக்காட்டுகள்

வாக்கியம்	w_{-2}	w_{-1}	w_{+1}	w_{+2}	பொருண்மை அடையாளம்
(e)	-	Determiner	Verb	Adj	FINANCE
(f)	Preposition	Determiner	Preposition	Determiner	SHORE

அட்டவணை 2: சொல் சூழல்களின் வேறுபட்ட அளவுகள்

சூழல் வடிவ அளவு	சூழல் எடுத்துக்காட்டுகள்
மானோகிராம் (monogram)	…bar…
பைகிராம் (bigram)	…friendly *bar*… …*bar* and…
டிரைகிராம் (trigram)	…friendly *bar* and… …*bar* and a… …and friendly bar…
சாளரம் (வடிவளவு ± n) (2n + 1) -grams	…warm and friendly bar and a cheerful… (n=3) …area, a warm and friencly bar and a cheerful dining roon… (n=5)
வாக்கியம்	There is a lounge area, a warm and friendly bar and a cheerful dining room.
	This is a very nice hotel. There is a lounge area, a warm and friendly bar and a cheerful dining room. A buffet style breakfast is

	served in the dining room between 7 A.M. and 10 A.M.

(e) The bank cashed my check, and

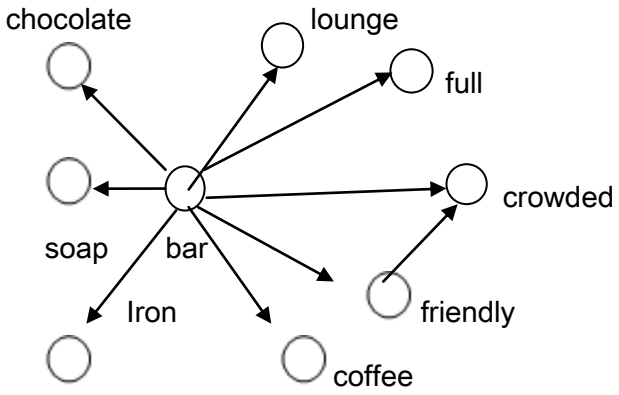
(f) We sat along the bank of the Tevere river,

இங்கு bank நமது இலக்குச் சொல் ஆகும்; மேலும் நம்முடைய திசையன்கள்/வெக்டர்கள் bank-இன் இடதுபக்கமும் வலதுபக்கமும் இரு சொற்களின் சொல்வகைப்பாடு அடையாளங்களுக்கு நான்கு வட்டாரப் பண்புக்கூறுகளையும் ஒரு பொருண்மை வகைப்படுத்தலையும் உட்படுத்தும் (நமது எடுத்துக்காட்டில் FINANCE அல்லது SHORE).

அட்டவணை 2-இல் bank என்ற சொல்லை இலக்காகக் கொண்டு வேறுபட்ட சூழல் வடிவ அளவுகளின் எடுத்துக்காட்டுகள் எடுத்துரைக்கப்பட்டுள்ளன. வடிவ அளவுகள் n-கிராம்களிலிருந்து, (அதாவது இலக்குச் சொல்லையும் உள்ளடக்கிய n சொற்களின் ஒரு தொடர்வரிசை) குறிப்பாக மானோகிராம் (n=1), பைகிராம் (n=2), டிரைகிராம் (n=3) என்பதிலிருந்து இலக்குச் சொல்லை உள்ளடக்கிய ஒரு முழு வாக்கியம் அல்லது பத்தி வரை வரிசைப்படுத்தும். சுற்றி இருக்கும் சொற்களின் இருப்பிடத்தின் அடிப்படையில் (இலக்குச் சொல்லின் வலதுபக்கதிலோ இடது பக்கத்திலோ) n-கிராமுகளுக்குப் பல விருப்பத்தேர்வுகளைச் செய்யவியலும் என்பதைக் கவனிக்கவும்; இருப்பினும் ±n வடிவ அளவுள்ள ஒரு சாளரம் இலக்குச் சொல்லைச் சுற்றி மையமிட்டுள்ள ஒரு (2n+1) கிராமாகும்.

கிளையமைப்புகள் அல்லது வரைபடங்கள் போன்ற கூடுதல் அமைப்பாக்கம் செய்யப்பட்ட உருப்படுத்தங்களைச் சொல்லின் சூழல்களை உருப்படுத்தம் செய்ய பயன்படுதவியலும்; இது ஆற்றலுடன் ஒரு முழுப் பனுவலையும் அளக்கவியலும். வெரோனிஸ் (Vèronis) சேர்ந்துவருகை வரைபடங்களை (coocurrence graphs) உருவாக்குகின்றார் (ஒரு எடுத்துக்காட்டு படம் 7-காட்டப்பட்டுள்ளது); மிஹல்செ மற்றும் பிறர் (Mihalcea et al 2004) மற்றும் நவிக்லி மற்றும் வெலர்டி (Navigli and Velardi 2005) பாதை (path), இணைப்பு ஆய்வு (link analysis) போன்றவைக்குப் பொருண்மையியல் வரைபடங்களை (semantic graphs) உருவாக்குகின்றனர்.

படம் 7: bar_n-இன் சாத்தியமான வரைபட உருப்படுத்தம்



சூழல் வெக்டர் (context vector) போன்ற தட்டையான உருப்படுத்தங்கள் (flat representations) கண்காணிக்கப்பட்ட பொருண்மைமயக்கநீக்க நெறிமுறைகளுக்குப் (supervised disambiguation methods) பொருத்தமானதாகும்; பயிற்சி மாதிரிகள் பொரும்பாலும் இவ்வடிவில் தான் உள்ளன. இதற்கு முரணாக அமைப்பாக்கம் செய்யப்பட்ட உருப்படுத்தங்கள் (unstructured representations) கண்காணிக்கப்படாத (unsupervised) மற்றும் அறிவு அடிப்படையிலான (knowledge based) நெறிமுறைகளில் கூடுதல் பயன்படுத்தப்படுகின்றன; அவைகள் பொருண்மைசார் வலையமப்புகளிலும் கணினிசார் சொற்களஞ்சியங்களிலும் குறியனாக்கம் செயப்பட்ட கருத்துருக்களுக்கு இடையிலான சொல்சார் மற்றும் பொருண்மைசார் பரஸ்பர உறவுகளை முழுவதுமாகப் பயன்படுத்தவியலும். (அமைப்பாக்கம் செய்யப்பட்ட மற்றும் அமைப்பாக்கம் செய்யப்படாத இரண்டிலும்) சூழலின் பொருத்தமான வடிவ அளவைத் தேர்ந்தெடுப்பது சொற்பொருண்மை மயக்கநீக்க வழிமுறை வரைவின் முன்னேற்றத்தில் ஒரு முக்கியக் காரணியாகும் என்பதைக் கவனத்தில் கொள்ளவும்; இது மயக்கநீக்கச் செயன்மையைப் பாதிப்பதாக அறியப்படுகின்றது [பார்க்க எடுத்துக்காட்டு Yarowsky and Florian 2002); Cuadros and Rogau (2006)].

பின்வரும் துணைப்பகுதியில், சொற் சூழல்களின் உருப்படுத்தத்திற்குப் பயன்படுத்தவியலும் வேறுபட்ட வகைபடுத்தும் நெறிமுறைகள் விளக்கப்பட்டுள்ளன.

4.1.4 வகைபடுத்தும் நெறிமுறைகளின் விருப்பத்தேர்வு (choice of classification methods)

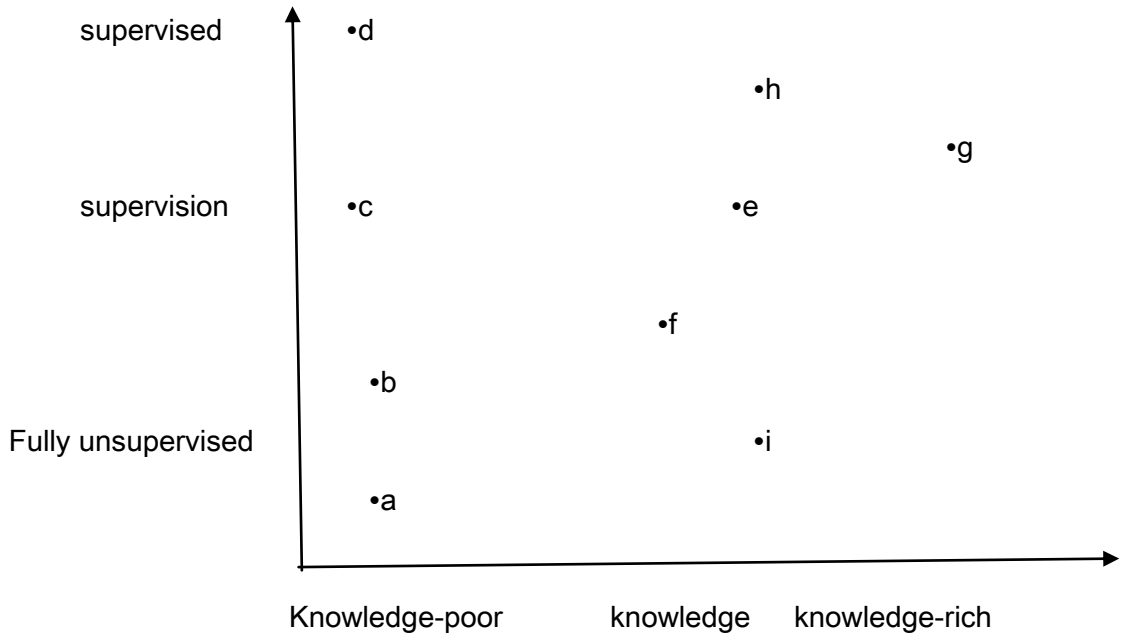
இறுதி நடவடிக்கை வகைப்படுத்தும் நெறிமுறையின் விருப்பத்தேர்வாகும். சொல் மயக்கநீக்கத்தின் தீர்வுக்கான பெரும்பான்மையான அணுகுமுறைகள் இயந்திரக் கற்றலின்

(machine learning) களத்திலிருந்து வளர்ந்ததாகும்/உருவானதாகும். இது வலுவான கண்காணிப்பிலிருந்து இக்களத்தின் ஒர் ஆழமான கையாளலுக்கு வேண்டித் தொடரியல் மற்றும் அமைப்பியல் அமைப்பொழுங்கு புரிதல் (syntactic and structural pattern recognition) அணுகுமுறைகள் (பார்க்க Mitchell 1997, Alpaydin 2004) வரைப் பரந்துகிடக்கின்றது. பரந்தநிலையில் சொற்பொருண்மை மயக்கநீக்கத்தின் இரு முக்கியமான அணுகுமுறைகளை வேறுபடுத்தலாம்:

கண்காணிக்கப்பட்ட சொற்பொருண்மை மயக்கநீக்கம் (Supervised WSD): இவ்வணுகுமுறைகள் புலக்குறிப்பு செய்யப்பட்ட பயிற்சிக் குழுமங்களிலிருந்து (labeled training sets) ஒரு வகைப்படுத்துவானை (classifer) கற்க இயந்திரம்-கற்றல் உபாயங்களை/நுட்பங்களை (machine learning techniques) அதாவது பொருத்தமான பொருண்மைப் புலக்குறிப்புகளுடன் (அல்லது வகுப்புகளுடன்) பல பண்புக்கூறுகள் அடிப்படையில் குறியனாக்கம் செய்யப்பட்ட எடுத்துக்காட்டுக்களின் குழுமங்களைப் பயன்படுத்துகின்றன.

கண்காணிக்கப்படாத சொற்பொருண்மை மயக்கநீக்கம் (unsupervised WSD): இந்நெறிமுறைகள் புலக்குறிப்பு செய்யப்படாத தரவுத்தொகுதிகள் அடிப்படையில் அமைந்தவை; சூழலில் ஒரு சொல்லின் பொருண்மை விருப்பத்தேர்வைத் தரவேண்டி கையால் பொருண்மை-அடையாளப்படுத்தப்பட்ட (sense tagged) எந்தத் தரவுத்தொகுதியையும் பயன்படுத்துவதில்லை.

மேலும் நாம் அறிவு-அடிப்படையிலான (Knowledge-based) (அல்லது அறிவுச்-செழுமையான, அல்லது அகராதி-அடிப்படையிலான) மற்றும் தரவுத்தொகுதி-அடிப்படையிலான (corpus-based) (அல்லது அறிவு-ஏழ்மையான) அணுகுமுறைகளுக்குள் வேற்றுமைகாணவேண்டும். முந்தையது இயந்திரம் படிக்கவியலும் அகராதிகள் (machine redable dictionaries), பொருட்புல அகராதிகள் (thesauries), மூலப்பொருண்மை ஆய்வுகள் (ontologies) போன்ற புறஞ்சார்ந்த சொல் மூலவளங்களைச் (external lexical resources) சார்ந்துள்ளது; பிந்தையது பொருண்மை மயக்கநீக்கத்திற்கு இம்மூலவளங்களில் எதையும் பயன்படுத்தவில்லை.



படம் 8-இல் சொற்பொருண்மை மயக்கநீக்க அணுகுமுறைகள் இருபரிமாண வெளியில் எடுத்துக்காட்டப்பட்டுள்ளன. வகைப்படுத்தல் கண்காணிப்பின் அளபு ஆகும்; அதாவது ஒழுங்குமுறையால் பயன்பத்தப்படும் பொருண்மை-அடையாளப்படுத்தப்பட்ட தரவுக்கும் புலக்குறிப்பு அடையாளப்படுத்தப்பட்ட தரவுக்கும் உள்ள விகிதமாகும்: ஒரு ஒழுங்குமுறை பொருண்மை-புலக்குறிப்புசெய்யப்பட்ட பயிற்சித் தரவை (training data) மிகக்கூடுதலாகப் பயன்படுத்தினால் அவ்வொழுங்குமுறை முழுவதும் (அல்லது வலுவாக) கண்காணிக்கப்பட்ட ஒழுங்குமுறை என்றும், ஒழுங்குமுறை ஒரு வகைப்படுத்துவானைக் (classifer) கற்க வேறுபட்ட விகிதங்களில் பொருண்மை-புலக்குறிப்புசெய்யப்பட்ட மற்றும் புலக்குறிப்பு செய்யப்படாத தரவைப் பயன்படுத்தினால் அவ்வொழுங்குமுறை பகுதி கண்காணிக்கப்பட்ட அல்லது வலுவற்ற (அல்லது மிகக்குறைவாக) கண்காணிக்கப்பட்ட ஒழுங்குமுறை என்றும் வரையறை விளக்கம்

செய்யப்படும். படத்தில் சமதளத்தின் பிரிவானது ஒழுங்குமுறையால் பயன்படுத்தப்படுகின்ற அகராதி வரையறை விளக்கங்கள், சொல்-பொருண்மைசார் உறவுகள், பொருட்புலப் புலக்குறிபுகள் மற்றும் பிறவற்றை உள்ளடக்கிய எல்லாப் பிற தரவுகளையும் கருத்தில்கொண்ட அறிவின் அளவை உருப்படுத்தம் செய்கின்றது.

துரதிருஷ்ட வசமாக, அடுத்தப் பகுதியில் விளக்கப்பட்டுள்ள சமதளத் திட்டவட்டமான நெறிமுறைகளை நிலையில் வைக்க இயலாது; ஏனென்றால் ஒரு வேவேறான எண்ணாக அறிவு மற்றும் கண்காணிப்பு என்பதன் அளவை மதிப்பிடுவதில் சிக்கல் இருக்கின்றது. இருப்பினும் நாம் (a)-இலிருந்து (i) வரை எழுத்துக்களுடன் படம் 8-இல் எடுத்துக்காட்டப்பட்டுள்ள இடத்தின் மீது பொதுவான அணுகுமுறைகளின் தோராயமான நிலையை அடையாளம் காண முயல்கின்றோம்: (a) எந்த அளவு அறிவையும் (பொருண்மை தெரிவடைவுகளைக் கூட) பயன்படுத்தாத முழுவதும் கண்காணிக்கப்படாத நெறிமுறைகள்; (b) & (c) கண்காணிப்பின் குறைந்த அளவு அல்லது பகுதி அளவை வேண்டும் மிகக் குறைவாக கண்காணிக்கப்பட்ட மற்றும் பகுதி கண்காணிக்கப்பட்ட அணுகுமுறைகள்; (d) கண்காணிக்கப்பட்ட அணுகுமுறைகள் (இயந்திரம்-கற்றல் வகைப்படுத்திகள்). இடத்தில் குறிப்பிட்ட பிற புள்ளிகளைத் தொடர்புபடுத்தும் அணுகுமுறைகள் கூடுதல் கடினமாகும் மற்றும் தனிநிலை நெறிமுறைகளின் குறிப்பிட்ட நிறைவேற்றலைச் சார்ந்து அமையும். இருப்பினும் வரைபட அமைப்பு (graph structutre) பொருண்மையில் வலையமைப்புகள் (semantic networks) போன்ற அமைப்புசார் சிறப்பியல்புகளைச் (g) சார்ந்திருக்கும் பெரும்பான்மையான அறிவு-அடிப்படையிலான அணுகுமுறைகள் பொதுவாக பொருள்விளக்க மேலுறல் (gloss overlap) (e) அல்லது சொற் பொருண்மை அதிகாரத்தை (sense dominance) (f) நிர்ணயிக்கும் நெறிமுறைகள் அடிப்படையிலான நெறிமுறைகளை விடக் கூடுதல் கண்காணிப்பையும் அறிவையும் பயன்படுத்தும். இறுதியாகப் கையால் குறியனாக்கம் செய்யப்பட்ட களப் புலக்குறிப்புகளைப் பயன்படுத்தும் பொருட்புலத்தால்-இயக்கப்பட்ட (domain-deiven) அணுகுமுறைகள் பொருண்மைச் சாத்தியங்களை மதிப்பிடுவதற்குக் கண்காணிக்கப்பட்ட கூறுகளை உட்படுத்தினால் அதை (h) என்ற புள்ளியைச் சுற்றி வைக்கவியலும்; அப்படியில்லை என்றால் (i) என்ற புள்ளியைச் சுற்றி வைக்கவியலும்.

இறுதியாக நாம் சொற்பொருண்மை மயக்கநீக்கத்தை அடையாளக்குறி/சொல்லலகு-அடிப்படையில்லனது (token-based) மற்றும் வகை-அடிப்படையிலானது (type-based) என்று

வகைப்படுத்தலாம். சொல்லலகு/அடையாளக்குறி அடிப்படையிலான அணுகுமுறைகள் ஒரு சொல் அது தோன்றும் சூழலைப் பொறுத்து அச்சொல்லின் ஒவ்வொரு நேர்வுடனும் ஒரு திட்டவட்டமான அர்த்தத்தைத் தொடர்புபடுத்துகின்றது. இதற்கு முரணாக வகை அடிப்படையிலான பொருண்மை மயக்கநீக்கம் ஒரு பனுவலுக்குள் ஒரு சொல்லானது ஒருமுகப்போக்காக ஒரே அர்த்தத்துடன் குறிப்பிடப்படுகின்றது என்ற ஊகம் அடிப்படையில் அமைகின்றது. இதன் விளைவாக இந்நெறிமுறைகள் முழுப் பனுவலின் ஆய்விலிருந்து ஒரு சொல்லுக்கு (முக்கியமான பொருண்மை என்று அழைக்கப்படுகின்ற) ஒரு பொருண்மையை அனுமானிக்கவும் அப்பனுவலுக்குள் ஒவ்வொரு நேர்வுக்கும் அதை ஒதுக்கவும் செய்யும். ஒரு சொல்லின் ஒவ்வொரு நேர்வுக்கும் பனுவல் முழுவதற்கும் பெரும்பான்மையான பொருண்மையை ஒதுக்கி அடையாளக்குறி/சொல்லலகு அடிப்படையிலான அணுகுமுறைகளை வகை அடிப்படையிலான முறையில் நிறைவேற்ற வேண்டி மாற்றியமைக்க இயலும் என்பதைக் கவனத்தில்கொள்ளவும்.

முதலாவது நாம் சொற்பொருண்மை மயக்கநீக்கத்திற்கு வேண்டிச் சுத்தமான கண்காணிக்கப்பட்ட மற்றும் கண்காணிக்கப்படாத அணுகுமுறைகளை மேலோட்டமாய்ப் பார்க்கின்றோம் (முறையே பகுதி 4.3 மற்றும் பகுதி 4.4). அடுத்தபடியாக சொற்பொருண்மை மயக்கநீக்கத்திற்கு அறிவு அடிப்படையிலான அணுகுமுறைகளை விளக்குகின்றோம் (பகுதி 4.5); மேலும் கலப்பின நெறிமுறைகளை முன்வைக்கின்றோம். இவ்வணுகு முறைகளின் பலவற்றிற்கு எடுத்துரைக்கப்பட்ட நிறைவேற்றம் உள்ளுக்குள்ளான அல்லது சிறிய அளவிலான பரிசோதனைகள் அடிப்படையிலானது. நாம் பகுதி 4.7-இலும் 4.8-இலும் பரிசோதனை மதிப்பீட்டில் கவனக்குவிப்பு செய்வோம்; இங்கு எவ்வாறு பொரும்பாலான சொற்பொருண்மை மயக்கநீக்க ஒழுங்கமைப்புகள் அவற்றின் நிறைவேற்றத்தை மேம்படுத்த வேண்டி உக்திகளின்/நுட்பங்களின் கலவையைத் தற்போது பயன்படுத்துகின்றன என்று பார்ப்போம்.

4.2 கண்காணிக்கப்பட்ட பொண்மைமயக்கநீக்கம் (Supervised Disambiguation)

கழிந்த 15 வருடங்களாக, இயற்கை மொழி ஆய்வு கையால் உருவாக்கப்பட்ட ஒழுங்குமுறைகளிலிருந்து (manually grafted system) தானியக்க வகைப்படுத்தும் நெறிமுறைகளின் (automatic classification methods) பயன்பாட்டிற்கு மாறியது (Cardie and Mooney 1999). இயந்திரம் கற்றல் நுட்பங்களை (machine learning techniques)) நோக்கிய

ஆர்வத்தின் அதிகரிப்பு சொற்பொருண்மை மயக்கநீக்கத்தின் சிக்கலுக்குப் பயன்படுத்தப்படும் பல கண்காணிக்கப்பட்ட அணுகுமுறைகளில் பிரதிபலித்தது. கண்காணிக்கப்பட்ட சொற்பொருண்மை மயக்கநீக்கம் கையால் பொருண்மை-அடையாளப்படுத்தப்பட்ட தரவுக் குழுமங்களிலிருந்து வகைபடுத்தியைத் தூண்ட இயந்திரம்-கற்றல் நுட்பங்களைப் பயன்படுத்துகின்றது. பொதுவாக [பெரும்பான்மையும் சொல் வல்லுநர் (word expert) என்று அழைக்கப்படுகின்ற] வகைப்படுத்திகள் (classifiers) ஒரு தனிச் சொல்லுடன் சம்பந்தமுள்ளது; மேலும் அச்சொல்லின் ஒவ்வொரு நேர்வுக்கும் தகுந்த பொருண்மையை ஒதுக்குவதற்கு வேண்டி ஒரு வகைப்படுத்தும் செயல்பாட்டை நிறைவேற்றுகின்றது. வகைப்படுத்தியைக் கற்கப் பயன்படுத்தப்படும் பயிற்சிக் குழுமம் (training set) எடுத்துக்காட்டுகளின் ஒரு குழுமத்தைக் சிறப்பம்சமாகக் கொண்டிருக்கும்; இதில் தரப்பட்ட இலக்குச் சொல் ஒரு நோக்கீட்டு அகராதியின் பொருண்மைத் தெரிவடைவிலிந்து ஒரு பொருண்மையால் கையால் அடையாளப்படுத்தப்பட்டுள்ளது.

பொதுவாகப் சொற்பொருண்மை மயக்கத்திற்குக் கண்காணிக்கப்பட்ட அணுகுமுறைகள் கண்காணிக்கப்படாத நெறிமுறைகளைக் காட்டிலும் நல்ல முடிவுகளைப் பெற்றுள்ளன. அடுத்த துணைப்பகுதிகளில், மிக அறிமுகமான இயந்திரம் கற்றல் நெறிமுறைகள் மதிப்புரை செய்யப்படும்; மேலும் சொற்பொருண்மை மயக்கநீக்கக் களத்தில் அவை சூழ்நிலயாக்கம் செய்யப்படும்.

4.2.1 தீர்மானப் பட்டியல்கள் (Decision lists)

சோதனை நேர்வுகளை வகைப்பாடு செய்வதற்குப் பயன்படும் வரிசைப்படுத்தப்பட்ட விதிகளின் குழுமம் (ordered set of rules) ஒரு தீர்மானப் பட்டியல் (decision list) (Rivest 1987) ஆகும் (சொற்பொருண்மை மயக்கநீக்க நிகழ்வைப் பொறுத்தவரையில் இது ஒரு இலக்குச் சொல்லுக்குப் பொருத்தமான பொருண்மையை ஒத்துக்கப் பயன்படும்.) இதை ''இல்லையென்றால்-மற்றொன்று/'if-then-else' விதிகளின் ஒரு பட்டியலாகக் காணவியலும். ஒரு பயிற்சிக் குழுமம் ஒரு குழுமப் பண்புக்கூறுகளைத் தூண்டுவதற்குப் பயன்படுத்தப்படுகின்றது. இதன் விளைவாக வகை விதிகள் (rules of the kind) (பண்புக்கூறு-மதிப்பு, பொருண்மை, மதிப்பெண்) உருவாக்கப்படுகின்றன. இவ்விதிகளின் இறங்குவரிசை மதிப்பெண் அடிப்படையில் இவ்விதிகளின் நிரல் (ordering of these rules) தீர்மானப் பட்டியலைக் கொண்டிருக்கும்.

w என்ற சொல்லின் நேர்வும் ஒரு பண்புக்கூறு வெக்டாராக இதன் உருப்படுத்தமும் தரப்படுகையில் தீர்மானப் பட்டியல் தணிக்கை செய்யப்படுகின்றது (பரிசோதிக்கப்படுகின்றது); உள்ளீட்டு வெக்டாருடன் பொருந்தும் உயர்ந்த மதிப்பெண் ஒதுக்க/தரப்பட வேண்டிய சொற்பொருண்மையைத் தெரிவு செய்கின்றது:

$$\hat{S} = \text{argmax}_{S_i \in Senses_{D(w)}} \text{score}(S_i).$$

யரொவ்ஸ்கியின் கருத்துப்படி (Yarowsky 1994), S_i என்ற பொருண்மையின் மதிப்பெண் (score) பண்புக்கூறு மதிப்பெண்களுக்கிடையில் மிகக்கூடுதலாகக் கணக்கிடப்படுகின்றது; இதில் பண்புக்கூறு f-இன் மதிப்பெண், பொருண்மை S_i-இன் நிகழ்வுத்தகமைகளின் லாகிரதம் தரப்பட்ட பண்புக்கூறு f-இன் பிற அர்த்தங்களின் நிகழ்வுத்தகமைகளின் மொத்தத்தால் வகுக்கப்படும் லாகிரதம் ஆகக் கணிக்கப்படுகின்றது.

$$\text{Score}(S_i) = \max \log \left(\frac{P(S_i \mid f)}{\sum_{j \neq i} P(S_j \mid f)} \right)$$

தீர்மானப் பட்டியலின் எளிமையான எடுத்துக்காட்டு அட்டவணை 3-இல் எடுத்துரைக்கப்பட்டுள்ளது. எடுத்துக்காட்டில் உள்ள முதல் விதி *bank* என்பதன் நிதிப் பொருண்மைக்குப் பயன்படுத்தப்படுகின்றது மற்றும் இடதுபுறச் சூழலாக *accounts with* என்பதை எதிர்பார்க்கின்றது; மூன்றாவது *bank* என்பதற்கு வினியோகப் பொருண்மையில் (எ.கா. *a bank of blood, a bank of food*) பயன்படுத்தப்படுகின்றது; இவ்வாறு சென்றுகொண்டிருக்கும்.

அட்டவணை 3: தீர்மானப் பட்டியலின் ஒரு எடுத்துக்காட்டு

Feature	Prediction	Score
Account with bank	Bank/FINANCE	4.83
Stand/V on/P…bank	Bank/FINANCE	3.35
Bank of blood	Bank/SUPPLY	2.48
work/V…bank	Bank/FINANCE	2.33
the left/J bank	Bank/RIVER	1.12
of the bank	-	0.01

மூல வெளிப்பாட்டின்/கருத்தாக்கத்தின் (Rivest 1987) தீர்மானத்தில் உள்ள ஒவ்வொரு விதியும் மதிப்பிடப்படாதது மற்றும் அவை பண்புக்கூறுகளின் இணைப்பைக் கொண்டிருக்கும்; ஆனால் யரோவ்ஸ்கியின் அணுகுமுறையில் ஒவ்வொரு விதியும் மதிப்பிடப்பட்டது மற்றும் அவை ஒரே ஒரு பண்புகூறை மட்டும் கொண்டிருக்கும். தீர்மானப் பட்டியல் முதல் சென்ஸ்வல் மதிப்பீட்டுப் போட்டிகளில் (Senseval evaluation competitions) (எ.கா. Yarowsky 2000) மிக வெற்றிகரமான நுட்பமாக அமைந்தது. அகிரெ மற்றும் மார்டினெஸ் (Agirre and Matinez 2000) கைகளால் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளின் குறைவால் ஏற்பட்ட அறிவுப்பேறு நெருக்கடியைத் (knowledge acquisition bottleneck) தணிக்கும் முயற்சியில் அவற்றைப் பயன்படுத்துகின்றனர்.

4.2.2 தீர்மானக் கிளைகள் (decision trees)

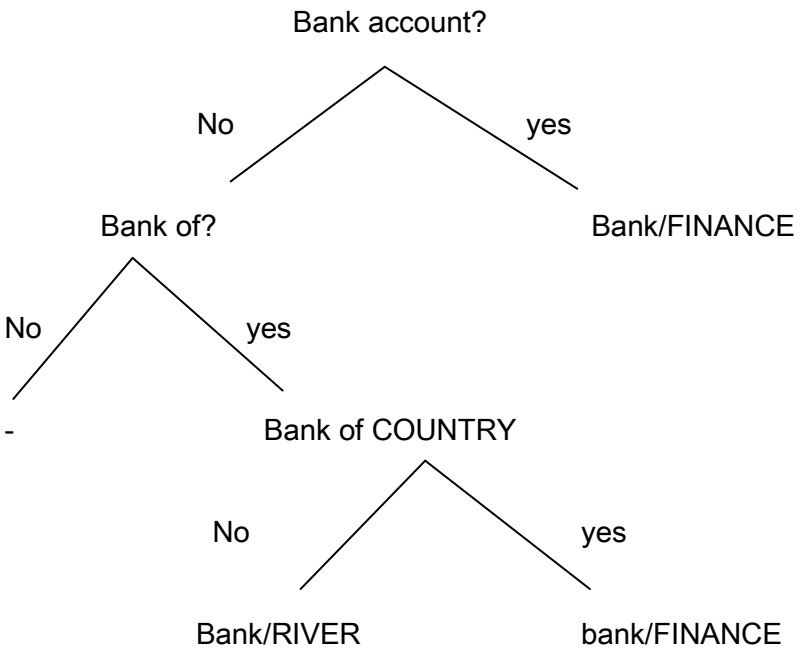
பயிற்சித் தரவுக் குழுமத்தைத் திரும்பத்திரும்பப் (மறுதரவாகப்) பிரிக்கும் ஒரு கிளை அமைப்புடன் வகைப்படுத்தும் விதிகளை உருப்படுத்தம் செய்யப் பயன்படுத்தப்படும் ஒரு பயனிலைசார் மாதிரி (predicative model) ஒரு தீர்மானக் கிளையாகும். தீர்மானக் கிளையின் ஒவ்வொரு உட்புறமான கணுவும் ஒரு பண்புக்கூறு மதிப்பின் மீதான ஒரு பரிசோதனையை உருப்படுத்தம் செய்யும். ஒரு இறுதிக் கணுவை (அதாவது ஒரு இலையை) அடைகையில் ஒரு வினைப்பாடு (predication) உருவாக்கப்படும்.

கழிந்த பத்தாண்டு காலங்களில் தீர்மானக் கிளைகள் சொற்பொருண்மை மயக்கநீக்கத்திற்கு அரிதாகப் பயன்படுத்தப்பட்டு வந்தது. தீர்மானக் கிளைகள் ஒழுங்காகவும் மனிதரால் படிக்கவியலும் வழியிலும் பயனிலைசார் மாதிரியை உருப்படுத்தம் செய்தாலும் அவை பல விவாதங்களால் பாதிக்கப்பட்டுள்ளன: பல எண்ணிக்கையிலான மதிப்புகளுடன் கூடிய பண்புக்கூறுகளின் காரணமாகத் தரவுப் பற்றாக்குறை, சிறிய அளவிலான பயிற்சி குழுமங்கள் காரணமாக வினைப்பாடுகளின் நம்பத்தகாமை.

படம் 9-இல் சொற்பொருண்மை மயக்கநீகத்திற்கு ஒரு தீர்மானக் கிளையின் ஒரு எடுத்துக்காட்டு எடுத்துரைக்கப்பட்டுள்ளது. எடுத்துக்காட்டாக “we sat on a bank of sand” என்ற வாக்கியத்தில் bank என்ற பெயர் வகைப்படுத்தப்படவேண்டுமென்றால், கிளை no-yes-no பாதையைப் பின்பற்றிய பின்னர் bank/RIVER என்ற பொருண்மையின் விருப்பத்தேர்வு

செய்யப்படுகின்றது. (-) என்ற பூஜிய மதிப்பு கொண்ட இலை குறிப்பிட்ட பண்புக்கூறு மதிப்புகள் அடிப்படையில் எந்த விருப்பத்தேர்வையும் செய்யவியலாது என்பதைக் சுட்டிக்காட்டுகிறது.

படம் 9: தீர்மானக் கிளையின் ஒரு எடுத்துக்காட்டு



4.2.3 நெய்வ் பெய்ஸ் (Naïve Bayes)

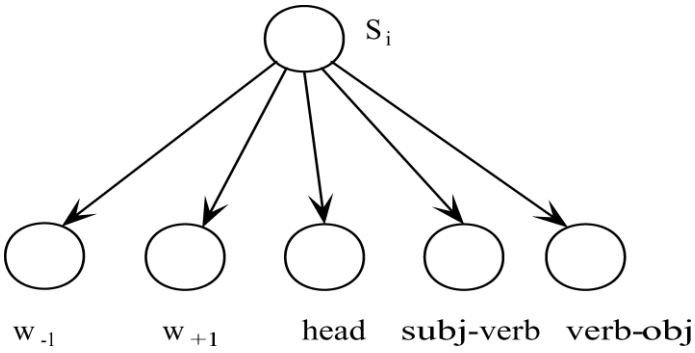
ஒரு நெய்வ் பெய்ஸ் வகைப்படுத்தி (Naïve Bayes classifier) பெய்ஸ்சின் வாய்பாட்டின்/சூத்திரத்தின் பயன்பாட்டின் அடிப்படியிலான ஒரு எளிய நிகழ்வுத்தகமைசார்/சாத்தியம்சார் வகைப்படுத்தியாகும் (probabilistic classifier). இது சூழலில் f_j பண்புகூறுகள் தரப்படுகையில் சொல் w-இன் ஒவ்வொரு அர்த்தம் S_i இன் கட்டுப்பாடு சாத்தியதின்/நிகழ்வுத்தகமையின் கணிப்பைச் சார்ந்திருக்கின்றது. பின்வரும் சூத்திரத்தை/வாய்ப்பாட்டை அதிகரிக்கும் $\hat{S}$ சூழலில் மிகப்பொருத்தமான அர்த்தமாகத் தெரிந்தெடுக்கப்படுகின்றது.

$$S = \underset{S_i \in SensesD(w)}{\operatorname{argmax}} P(S_i \mid f_1, \ldots, f_m) = \underset{S_i \in SensesD(w)}{\operatorname{argmax}} \frac{P(f_1, \ldots, f_m \mid S_i) P(S_i)}{P(f_1, \ldots, f_m)}$$

$$= \underset{S_i \in SensesD(w)}{\operatorname{argmax}} P(S_i) \prod_{J=1}^{m} P(f_j \mid S_i),$$

இதில் *m* என்பது பண்புக்கூறுகளின் எண்ணிக்கை, இறுதி வாய்ப்பாட்டு அர்த்தம் தரப்படுகையில், பண்புகூறுகள் கட்டுப்பாடு அடிப்படையில் சுதந்திரமானது என்ற அனுபவம்சாரா அனுமானத்தின் அடிப்படையில் கிடைக்கப்பெறுகின்றது. (வகுவெண் கணக்கீடு பாதிக்காத காரணத்தால் அது நீக்கப்படுகின்றது). முறையே $P(S_i)$ மற்றும் $P(f_j \mid S_i)$ என்ற நிகழ்வுத்தகமைகள் தீர்மானிக்கப்படுகின்றது. அர்த்தம் *Si*-இன் இருக்கையில் S_i மற்றும் பண்புக்கூறு f_j -இன் பயிற்சிக் குழுமத்தில் ஒப்பீட்டு நிகழ்வுஎண்ணிக்கைகளாக பூஜிய எண்ணிக்கைகள் நேர்செய்யப்படவேண்டும். எடுத்துக்காட்டாக அவற்றை $P(S_i)/N$ என்பதால் இடம்பெயர்க்க இயலும். இதில் *N* பயிற்சிக்குழுமத்தின் அளவாகும் [Ng 1997; Escudero et al. 2000c]. இருப்பினும் இத்தீர்வு மொத்தத்தின் நிகழ்தகமைகளை 1-க்கு அதிகமாகக்கொண்டு செல்கின்றது. பின்வலிப்பு (backoff) அல்லது சொருகுதல் (interpolation) நடவடிக்கைகள் இச்சிக்கல்களை விலக்குவதற்குப் பதிலாகப் பயன்படுப்பட இயலும்.

படம் 10: பெய்சியன் வலையமைப்பு மாதிரியின் எடுத்துக்காட்டு



படம் 10-இல் நெய்வ் பெய்சின் வலையமைப்பின் (naïve Bayesian network), எடுத்துக்காட்டாகப் பின்வரும் பண்புகூறுகள் தரப்படுகையில் The bank cashed my check என்ற வாக்கியத்தில் bank என்ற பெயர்சொல்லின் நேர்வை வகைப்படுத்த விரும்புவதாக வைத்துக்கொள்வோம்: {w_{-1} = *the*, w_{+1} = *cashed*, *head* = *bank*, *subj-verb* =*cash*, *verb-obj*= −}, இதில் இரண்டு பண்புக்கூறுகள் இலக்கு வாக்கியத்தில் பெயர்சொல் *bank*இன் இலக்கணப் பங்களிப்பை எழுவாயாகவும் நேரடிச்

செயப்படுபொருளாகவும் குறியனாக்கம் செய்கின்றது. *bank*இன் நிதி அர்த்தங்கள் தரப்படுகையில் நாம் பயிற்சிக் குழுமத்திலிருந்து இந்த ஐந்து பண்புக்கூறுகளின் நிகழ்வுத்தகமையை நிர்ணயிப்பதாக வைத்துக்கொள்வோம்: $P(w_{-1} = the \mid bank/\text{FINANCE}) = 0.66$, $P(w_{+1} = cashed \mid bank/\text{FINANCE}) = 0.35$, $P(head = bank \mid bank/\text{FINANCE}) = 0.76$, $P(subj\text{-}verb = cash \mid bank/\text{FINANCE}) = 0.44$, $P(verb\text{-}obj = - \mid bank/\text{FINANCE}) = 0.6$. மேலும் நாம் bankஇன் நேர்வின் நிகழ்வுத்தகமையைப் பின்வருமாறு நிர்ணயிக்கின்றோம்: $P(bank/\text{FINANCE}) = 0.36$. இறுதி மதிப்பெண் பின்வருவதாகும்:

$$score(bank/\text{FINANCE}) = 0.36 \cdot 0.66 \cdot 0.35 \cdot 0.76 \cdot 0.44 \cdot 0.6 = 0.016.$$

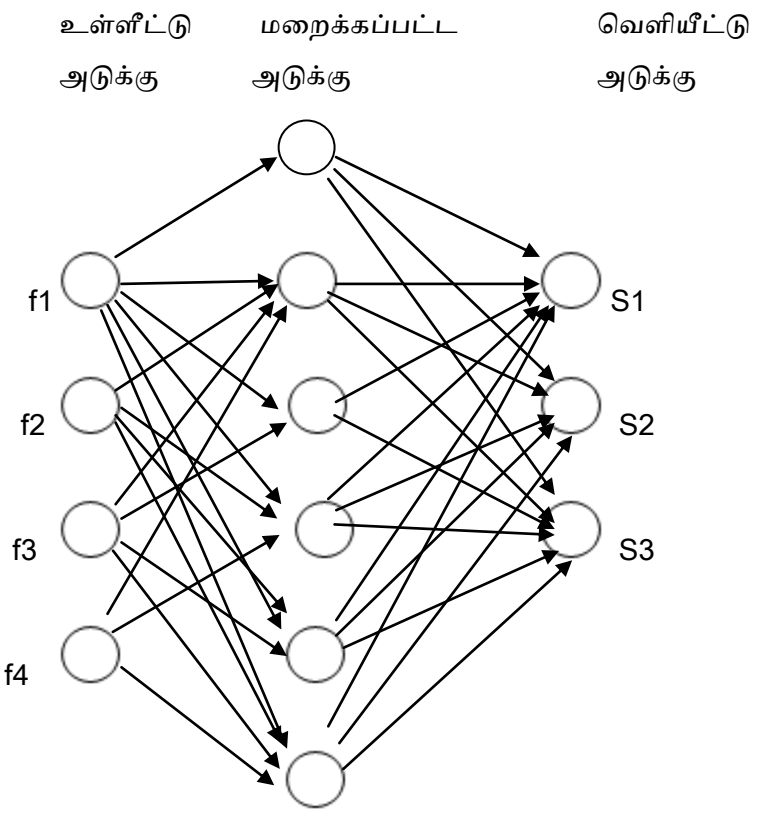
சுதந்திரமான அனுமானத்திற்குப் பிறகும், இந்நெறிமுறை பிற கண்காணிக்கப்பட்ட நெறிமுறைகளுடன் நன்றாக ஒப்புமை காட்டுகின்றது [Mooney 1996; Ng 1997; Leacock et al. 1998; Pedersen 1998; Bruce and Wiebe 1999].

4.2.4 நரம்புசார் வலையமைப்பு (neural network)

ஒரு நரம்புசார் வலையமைப்பு (neural network) (McCulloch and Pitts 1943) இணைப்பியலார் அணுகுமுறை (connectionsist approach) அடிப்படையிலான தரவு ஆய்வுக்கான ஒரு கணினிசார் மாதிரியை பயன்படுத்தும் செயற்கை நரம்பணுக்களின் ஒன்றோடொன்று இணைக்கப்பட்ட குழுவாகும் (interconnected group of artificial neurons). உள்ளீட்டுப் பண்புக்கூறு (input feature), விரும்பப்பட பதில்விளைவு (desired response) என்ற இணைகள் கற்றல் வழியமைப்புக்கு உள்ளீடாகும். இதன் நோக்கம் விரும்பப்பட்ட பதில்விளைவுகளுக்குப் பொருத்தமான மேலுறல் செய்யாத குழுமங்களாகப் பயிற்சிச் சூழல்களைப் பிரிப்பதற்கு உள்ளீட்டுப் பண்புக்கூறுகளைப் பயன்படுத்துவதாகும். புதிய இணைகள் தரப்படுவதால் இணைப்பு மதிப்புகள் விரும்பப்பட்ட பதில்விளைவை உருப்படுத்தம் செய்யும் வெளியீட்டு அலகு பிற வெளியீட்டு அலகைவிடக் கூடுதல் தீவிரத்தைக் (activation) கொண்டிருக்கும்படித் தொடர்ச்சியாக ஒழுங்குபடுத்தப்படுகின்றது. படம் 11-இல் ஒரு பன்னடுக்கு புலனுணர்வு நரம்புசார் வலையமைப்பின் (multilayered perception neural network) (ஒரு புலனுணர்வு என்பது முன்நோக்கு ஊட்டு (feedforward) நரம்பு வலையமைப்பின் எளிய வகையாகும்) ஒரு எடுத்துக்காட்டு எடுத்துரைக்கப்படுகின்றது; இது நான்கு பண்புக்கூறுகளின் மதிப்புகளால்

ஊட்டப்பட்டுள்ளது; அவை சூழலில் ஒரு இலக்குச் சொல்லின் முன்று பொருண்மைகளின் ஒன்றுக்கொன்று பொருத்தமான மதிப்பை (அதாவது மதிப்பெண்ணை) வெளியீடு செய்கின்றது.

படம் 11: நான்கு பண்புக்கூறுகளையும் மூன்று பதில்விளைவுகளையும் கொண்ட சொற்பொருண்மை மயக்கநீக்கத்திற்கு ஒரு முன்நோக்கு ஊட்டு நரம்பு வலைப்பின்னலின் ஒரு எடுத்துக்காட்டு.



நரம்புசார் வலைப்பின்னல்கள் விரும்பப்பட்ட பதில்விளைவின் பொருத்தமான அலகின் வெளியீடு ஒவ்வொரு பயிற்சி எடுத்துக்காட்டுக்கும் ஏதாவது பிற அலகின் வெளியீட்டைவிட பெரிதாக இருக்கும் வரை பயிற்சி செய்யப்படுகின்றது. மதிப்பீட்டிற்கு, வலையமைப்பினால் நிர்ணயிக்கப்பட்ட வகைப்படுத்தல் மிகப்பெரிய வெளியீடு கொண்ட அலகால் தரப்பட்டுகிறது. வலையமைப்பில் உள்ள எடைகள்/மதிப்புகள் (weights) நேர்நிலையாகவோ எதிர்நிலையாகவோ இருக்கலாம்; இவ்வாறு சான்றுகளின் திரட்டு ஒரு பொருண்மை விருப்பத்தேர்வுக்குச் சாதகமாகவோ எதிராகவோ இருப்பதைச் சாத்தியமாக்கும்.

பல ஆய்வுகளில் பிற கண்காணிக்கப்பட்ட நெறிமுறைகளை ஒப்பிடுகையில் நரம்புசார் வலையமைப்புகள் நன்றாகச் செயலாற்றுவதாகக் காட்டப்பட்டுள்ளது [Leacock et al. 1993; Towell and Voorhees 1998; Mooney 1996]. இருப்பினும் இப்பரிசோதனைகள் யாவும் பெரும்பான்மையும் ஒரு சிறிய எண்ணிக்கையிலான சொற்களின் மீதுதான் நடத்தப்பட்டுள்ளன. முடிவுகளைப் பொருள்கொள்வதில் உள்ள கடினம், மிகப்பெரிய அளவில் பயிற்சித் தரவின் தேவை, தொடக்கநிலை, நலிவு போன்ற காரணிகளைத் தக்கவாறு அமைத்தல் போன்றன நரம்புசார் வலையமைப்புகளின் முக்கியக் குறையாகக் கூறப்படுகின்றது.

4.2.5 முன்மாதிரி அடிப்படையிலான அல்லது எடுத்துக்காட்டு அடிபடையிலான கற்றல் (Examplar-based or Instance based learning)

முன்மாதிரி அடிப்படையிலான (exemplar-based) [(அல்லது எடுத்துக்காட்டு அடிபடையிலான (instance-based) அல்லது நினைவக அடிப்படியிலான (memory based)] கற்றல் ஒரு கண்காணிக்கப்பட்ட வழிமுறைவரைவாகும்; இதில் வகைப்படுத்தல் மாதிரி (classification model) எடுத்துக்காட்டுகளிலிருந்து உருவாக்கப்படுகின்றது. இம்மாதிரியானது பண்புக்கூறு வெளியில் (feature space) புள்ளிகளாக நினைவகத்தில் எடுத்துக்காட்டுகளைத் தக்கவைக்கின்றது; புதிய எடுத்துக்காட்டுகள் வகைப்படுத்தப்படும் போது அவை ஏறுமுகமாக மாதிரியில் சேர்க்கப்படுகின்றன.

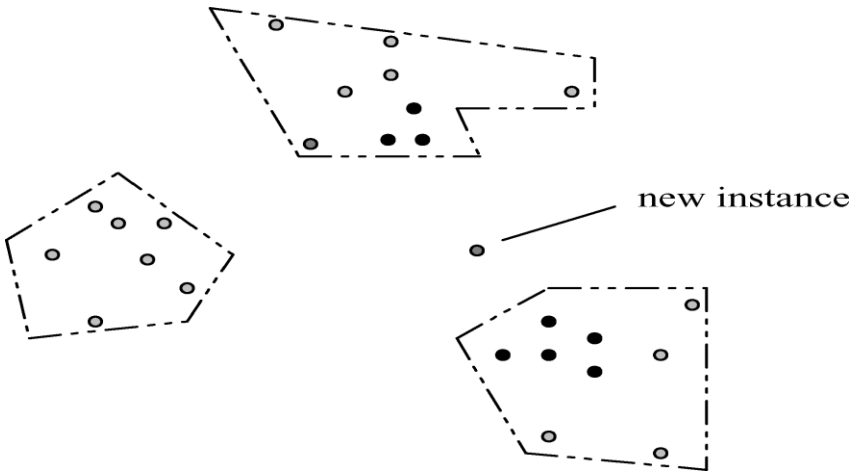
இப்பகுதியில் இவ்வகையில்படும் ஒரு குறிப்பிட்ட அணுகுமுறையான k-மிக அண்மைப்பட்ட அடுத்தவர் வழிமுறைவரைவைப் [k-nearest neighbor (kNN) algorithm] பார்ப்போம்; இது சொற்பொருண்மை மயக்கநீக்கத்தில் மிக உயர்வாகச் செயற்படும் நெறிமுறைகளில் ஒன்றாகும். [Ng 1997; Daelemans et al. 1999].

கே.என்.என்-இல் (kNN)/ மிக அண்மைப்பட்ட அடுத்தவரில் m பண்புக்கூறு மதிப்புகள் அடிப்படையில் உருப்படுத்தம் செய்யப்பட்ட புதிய எடுத்துக்காட்டின் வகைப்படுத்தல் (classification) $\mathbf{x} = (x_1, \ldots, x_m)$, முன்னர் சேகரித்து வைக்கப்பட்டுள்ள எடுத்துக்காட்டுடன் மிக ஒற்றுமையுள்ள அர்த்தம் அடிபடையில் அமைந்ததாகும். X-க்கும் ஒவ்வொரு சேகரித்து வைக்கப்பட்டுள்ள எடுத்துக்காட்டு $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ என்பதற்கும் இடையிலுள்ள தூரம் கணிக்கப்படுகின்றது, எடுத்துக்காட்டாகப் பின்வரும் ஹாமிங் (Hamming) தூரம்:

$$\Delta(\mathbf{x}, \mathbf{xi}) = \sum_{j=1}^{m} wj\ \delta\ x\ j\ ,\ xi\ j$$

இதில் *wj* என்பது *j* மற்றும் $\delta(x_j, x_{ij})$ is 0 if $x\ j = x_{i\ j}$ ஆவது பண்புக்கூறின் மதிபீடு/எடையாகும்; இல்லையென்றால் 1. அண்மைப்பட்ட நேர்வுகளின் குழுமம் k தெரிவுசெய்யப்படுகின்றது மற்றும் புதிய நேர்வுகள் குழுமத்திற்குள் பெரும்பாலான எண்ணிகையுள்ள நேர்வுகளுக்கு ஒதுக்கப்படுகின்ற வகுப்பைச் சார்ந்ததாக அனுமானிக்கப்படுகின்றது.

படம் 12: இருபரிமாண தளத்தின் மீது kNN வகைப்படுத்தலின் எடுத்துக்காட்டு



மிக அண்மைப்பட்ட அடுத்தவர்களின் எண் kஐ பரிசோதனை வாயிலாக நிர்ணயிக்கவியலும். படம் 12 காட்சியாக எவ்வாறு புதிய நேர்வுகள் k-ஆவதன் மிக அண்மைப்பட்ட அடுத்தவருடன் தொடர்புபடுத்துகின்றது என எடுத்துக்காட்டுகின்றது: ஒரே அர்த்தத்திற்கு ஒதுக்கப்படுகின்ற/ தரப்படுகின்ற நேர்வுகள் பல்கோணத்தில்/பாலிகானில் அடக்கப்பட்டுள்ளது; கறுப்புப் புள்ளிகள் புதிய நேர்வுகளின் மிக அண்மைப்பட்ட அடுத்தவர்கள்; பிற நேர்வுகள் சாம்பல் நிறத்தில் வரையப்பட்டுள்ளன. புதிய நேர்வு ஐந்து கறுப்புப் புள்ளிகள் உள்ள அடி வகுப்பிற்கு ஒதுக்கப்படுகின்றது.

டேல்மான்ஸ் மற்றும் பிறர் (Daelemens et al 1999) முன்மாதிரி அடிபடையில்லான அணுகுமுறைகளின் போக்கு மேம்பட்டதாக இருக்கின்றது என்று வாதிடுகின்றனர்; ஏனென்றால் அவை விதிவிலக்குகளைப் புறக்கணிப்பதிலை மற்றும் புதிய எடுத்துக்காட்டுகள் கிடைப்பதால்

பொருண்மை மயக்க நீக்கத்திற்காக மேலும் உதவிகளைத் குவிக்கின்றன. தற்போது முன்மாதிரி அடிப்படையிலான கற்றலானது சொற்பொருண்மை மயக்கநீக்கத்தில் புதிய கருத்துக்களையும் மிக தற்காலத்திற்கு ஏற்ற பண்புக்கூறுகளையும் உள்ளடக்கிய நிறைவேற்றத்தை (state-of-the-art performance) செயற்படுத்தியுள்ளது (Escudero et al 2000b; Fujii et al 1998; Ng and Lee 1996; Hoste et al 2002; Decadt et al 2004).

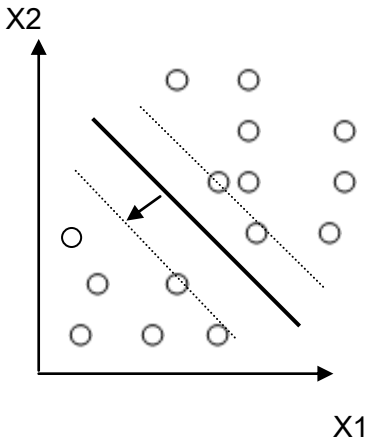
4.2.6 ஆதரவு திசையன் பொறிகள்/சப்போர்ட் வெக்டார் மெஷியன்கள் (Support Vector Machines (SVM)

போசர் மற்று பிறரால் (Boser et al 1992) அறிமுகப்படுத்தப்பட்ட இந்நெறிமுறை நேர்நிலை எடுத்துக்காட்டுகளை எதிர்நிலை எடுத்துக்காட்டுகளிலிருந்து பிரிக்கும் பயிற்சி குழுமத்திலிருந்து ஒரு கோடான உயர்சமதளத்தைக் (லீனியார் ஹைப்பர்பிளேனை (linear hyperplane)) கற்கும் கருத்தை அடிப்படையாகக் கொண்டதாகும். (ஆதரவு திசையன்கள்/ சப்பொர்ட் வெக்டார்கள் (support vectors) என்று அழைக்கப்படுகின்ற) உயர்சமதளம் (ஹைபர் பிளேன்) மிக அண்மையிலான நேர்நிலை மற்றும் எதிர்நிலை எடுத்துக்காட்டுகளின் தூரத்தை அதிகப்படுத்தும் உயர்வெளியிடத்தின் (ஹைபர்ஸ்பேஸ்சின் (hyperspace)) புள்ளியில் இடங் காணப்படுகின்றது. மற்றொருவிதமாகக் கூறினால் ஆதரவு திசையன் பொறிகள் (சப்போர்ட் வெக்டார் மெஷின்கள் (support vector mechines (SVMs)) அதே சமயத்தில் அனுபவம்சார் வகைப்படுத்தல் பிழையைக் குறைக்கவும் நேர்நிலை எடுத்துக்காட்டுகளுக்கும் எதிர்நிலை எடுத்துக்காட்டுகளுக்கும் இடையிலுள்ள வடிவயியல்சார் இடவெளியை (geometric margin) அதிகரிக்கவும் செய்யும்/முயலும். படம் 13 வடிவயியல்சார் உள்ளுணர்வை எடுத்துக்காட்டும்: கெட்டியாக இருக்கும் கோடு எடுத்துக்காட்டுகளின் இரண்டு வகுப்புகளைப் பிரிக்கும் சமதளத்தை உருப்படுத்தம் செய்கின்றது; இரண்டு புள்ளியிட்ட கோடுகள் மிகநெருங்கிய நேர்நிலை மற்றும் எதிர்நிலை எடுத்துக்காட்டுகளுக்கு டான்ஜென்டாக வரும் சமதளத்தைக் காட்டும். கோடான (லீனியார்) வகைபடுத்தி இரு தனிமங்களின் அடிப்படையிலானது: (பயிற்சிக் குழுமத்தை மதிப்பிடும் மற்றும் பண்புக்கூறுகளை உருப்படுத்தம் செய்யும் அவற்றின் கூறுகள்) உயர்சமதளத்திற்குச் (ஹைப்பர்பிளேனுக்கு) செங்குத்தான w என்ற ஒரு வெயிட் வெக்டார் மற்றும் மூலத்திலிருந்து ஹைப்பர்பிளேனின் ஆஃப்செட்டை நிர்ணயிக்கும் b என்ற பையாஸ். ஒரு

புலக்குறிப்பு செய்யப்படாத x என்ற எடுத்துக்காட்டு f (x) = w.x+b ≥ 0 என்றால் நேர்நிலையாக வகைப்படுத்தப்படும் (அல்லாவிடில் எதிர்நிலையாக வரும்).

உயர்சமதளம் (ஹைப்பர்பிளேன்) கோடாக இடத்தைப் பகுக்க இயலாது போகலாம். இந்நேர்வுகளில் பயிற்சிக் குழுமத்தைத் தக்கவாறு அமைத்துக்கொள்ள தளர்வான மாறிகளைப் பயன்படுத்தவும் இடப்பரப்பின் கோடுசார் பிரிவுக்கு அனுமதிக்கவும் இயலும்.

படம் 13: எஸ்.வி.எம்.-இன் வடிவவியல்சார்உள்ளுணர்வு



எஸ்.வி.எம்/SMV ஒரு ஈரிணையான வகைப்படுத்தியாக (binary classifier) இருப்பதால், சொற்பொருண்மை மயக்கநீக்கத்திற்குப் பயன்படவேண்டி அது பன்வகுப்பு வகைப்படுத்தலுக்கு (multiclass classification) (அதாவது ஒரு இலக்குச் சொல்லின் பொருண்மைகள்) மாற்றியமைக்கப்பட வேண்டும். எடுத்துக்காட்டாக, பன்வகுப்புச் சிக்கலைப் பல ஈரிணையான வகைப்படுத்தல்களாகச் சுருக்குவது ஒரு எளிய சாத்தியம் ஆகும்: பிற எல்லாப் பொருண்மைகளுக்கும் நேரெதிரான வகைப் பொருண்மை Si. இதன் விளைவாக மிக அதிக உறுதியுள்ள அர்த்தம் தெரிந்தெடுக்கப்படுகின்றது.

SVM-இன் வகைப்படுத்தல் வாய்ப்பாடை சப்போர்ட் வெக்டார்களின் செயல்பாடுகளாகச் சுருக்கவியலும் எனக்காட்டவியலும்; அது அதன் கோட்டு வடிவில் (linear form) வெக்டாரின் இணைகளின் புள்ளி விளைபொருளை (dot product) நிர்ணயம் செய்யும். கூடுதல் பொதுவாக, X மற்றும் Y என்ற இரண்டு வெக்டார்களுக்கிடையிலான ஒற்றுமை மூல இடப்பரப்பை (எ.கா.

பயிற்சி மற்றும் பரிசோதனை எடுத்துக்காட்டுகளின்) ஒரு பண்புக்கூறு இடப்பரப்புடன் பொருத்தும் கெர்னல் (kernel) என்ற செயல்பாட்டால் கணக்கிடப்படும். இதன்படி $k(x, y) = \Phi(x) \cdot \Phi(y)$, இதில் Φ ஒரு மாற்றமாகும் (மிக எளிதான கர்னல் புள்ளி விளைபொருள் $k(x, y) = x \cdot y$). இச்சிக்கலுக்குப் பொருத்தமான மூல உருப்படுத்ததை மாற்றுவதற்கு (கெர்னல் டிரிக்/kernel trick என்று அழைக்கப்படுகின்ற) ஒரு கோடற்ற மாற்றம் தெரிந்தெடுக்கப்படலாம் (non-linear transformation). அளகை திருப்பம் (parameter turning) அடிப்படையில் சந்தர்ப்பத்திற்குத் தக்கவாறு மாறுவதன் உயர்ந்த அளபுடன் கூடி கெர்னல் நெறிமுறைகளுடன் உயர்ந்த பரிமாணங்களுக்கு வெக்டார் இடங்களைப் பொருத்துவதற்கு இயலுவது எஸ்.வி.எம்.-இன் முக்கிய வெற்றிக்காரணிகளில் சிலவாகும்.

எஸ்.வி.எம்/SVM பனுவல் வகைப்பாடு செய்தல், பகுத்துகுறித்தல், சொற்பொருண்மை மயக்கநீக்கம் என்பனவறை உள்ளடக்கிய இயற்கை மொழி ஆய்வில் பல சிக்கல்களுக்குப் பயன்படுத்தப்படுகின்றது. SVM பல கண்காணிக்கப்பட்ட அணுகுமுறைகளுடன் ஒப்பிடுகையில் நல்ல முடிவுகளைத் தருவதாகக் காட்டப்பட்டுள்ளது.

4.2.7 ஒருங்கிணைந்த நெறிமுறைகள் (Ensemble Methods)

சில சமயங்களில் ஒட்டுமொத்தமான பொருண்மைமயக்கநீக்கத் துல்லியத்தை மேம்படுத்த ஒன்றுசேர்க்கப்பட வேண்டிய வேறுபட்ட வகைப்படுத்திகள் உள்ளன. ஒருங்கிணைந்த நெறிமுறைகள் என்று அழைக்கப்படுகின்ற ஒன்றிணைப்புத் திட்டங்கள்/நடவடிக்கைகள் வேறுபட்ட இயல்புகள் கொண்ட அதாவது தனித்தன்மையுடன் வேறுபட்ட இயல்புகள் உள்ள கற்றல் வழிமுறை வரைவுகள் ஒன்றுசேர்ந்தது. வேறுவிதமாகக் கூறினால் பண்புக்கூறுகள் தனிச்சிறப்பாக வேறுபட்ட, சுதந்திரச் சாத்தியமான பயிற்சித் தரவின் கருத்துக்களைத் தருமாறு விருப்பத்தேர்வு செய்யப்படவேண்டும் (எ.கா. சொல்சார், இலக்கணசார், பொருண்மையியல்சார் பண்புகூறுகள் போன்றன).

ஒருங்கிணைக்கப்பட்ட நெறிமுறைகள் அவை தனிநிலையான கண்காணிக்கப்பட்ட அணுகுமுறைகளின் வலுக்குறைவை விஞ்சுவதற்கு உதவுவதால் கூடுதல் பிரபலமானதாக மாறுகின்றது. சமீபகாலத்தில் மதிப்பீட்டு நடவைக்கைகளின் பங்கெடுக்கும் பல ஒழுங்குமுறைகள் இந்நெறிமுறைகளைப் பயன்படுத்துகின்றன. கிலென் மற்றும் பிறர் (Klein et al 2002) மற்றும் ஃப்ளோரியன் மற்றும் பிறர் (2002) கண்காணிக்கப்பட்ட சொற்பொருண்மை மயக்கநீக்கத்தின்

ஒருங்கிணைப்பை ஆய்ந்து Senseval-2 சொல்சார் மாதிரி வேலையில் புதிய கருத்துக்களையும் தற்காலத்துக்கேற்ற பண்புக்கூறுகளையும் உள்ளடக்கிய முடிவுகளை (state-of-art) பெற்றுள்ளனர். பராடி மற்றும் பிறர் (Brody et al. 2006) கண்காணிக்கப்படாத சொற்பொருண்மை மயக்கநீக்க நெறிமுறைகளின் ஒருங்கிணைந்த ஆய்வுகளை எடுத்துரைக்கின்றார். Senseval-3 எல்லாச்-சொற்கள் சொற்பொருண்மை மயக்கநீக்கச் (Senseval-3 all-words WSD) செயற்பாடுகள் போன்ற நிலைபேறுபெற்ற பரிசோனைக் குழுமத்தில் பயன்படுத்தப்படும்போது ஒருங்கிணைந்த நெறிமுறைகள் கண்காணிக்கப்படாத ஒழுங்குமுறைகளுக்குள் state-of-art நிறைவேற்றத்தை விஞ்சிநிற்கிறது.

தனி வகைப்படுத்திகளை (single classifiers) வேறுபட்ட நடவடிக்கைகளுடன் (different strategies) ஒருங்கிணைக்க இயலும்: இங்கு பெரும்பான்மை வாக்களிப்பு (majority voting), நிகழ்தகவுக் கலப்பு (probability mixture), தர-அடிப்படை ஒருங்கிணைப்பு (rank-based combination), அடாபூஸ்ட் (AdaBoost) என்பன அறிமுகப்படுத்தப்படுகின்றன. நிறை அறியப்பட்ட வாக்களிப்பு (weighted voting), மிக அதிக தேர்ந்தெடுக்கப்பட்ட தகவல் ஒருங்கிணைப்பு (maximum entropy combination) போன்ற பிற ஒருங்கிணைந்த நெறிமுறைகள் தொடர்புள்ள படைப்புகளில்/இலக்கியங்களில் ஆயப்பட்டுள்ளன. பின்வருவதில் முதல்-நிரல் வகைப்படுத்திகள் (அதாவது ஒருங்கிணைக்கப்படுவிருக்கும் ஒழுங்குமுறைகள், அல்லது ஒருங்கிணைந்த கூறுகள்) $C1, C2, \ldots, Cm$ எனக் குறிப்பிடப்படுகின்றது.

4.2.7.1 பெரும்பான்மை வாக்களிப்பு (Majority voting)

w என்ற இலக்குச் சொல் தரப்படுகையில் ஒவ்வொரு ஒருங்கிணைக்கப்பட்ட கூறும் w-இன் ஒரு அர்த்ததிற்கு ஒரு வாக்களிப்பு தர இயலும். அதிக வாக்களிப்பு உள்ள $\hat{S}$ என்ற அர்த்தம் தேர்ந்தெடுக்கப்படும்.

$$\hat{S} = \text{argmax}_{Si \in SensesD(w)} | \{ j : vote(C_j) = S_i \} |,$$

இதில், ஒரு வகைப்படுத்தி தரப்படுகையில் வாக்களிப்பு (vote) வகைப்படுத்தியால் தெரிந்தெடுக்கப்படும் அர்த்தத்தை வெளியீடு செய்யும் செயற்பாங்கு ஆகும். போட்டி நேர்வில் அதிக வாக்களிப்பு கொண்ட அர்த்தங்களுக்கு இடையில் சீரற்ற தெரிவைச் (random choice) செய்ய இயலும்.

4.2.7.2 நிகழ்தகமை கலப்பு (probability mixture)

முதல் நிரல் வகைப்படுத்தி w என்ற இலக்குச் சொல்லிலுக்கு அர்த்தங்களின் நம்பிக்கை மதிப்பெண்ணைத் (confidence score) தருவதாக எடுத்துக்கொண்டால் w-வின் அர்த்தத்தின் மீது விநியோக நிகழ்வு தகமைக்கு இம்மதிப்பெண்களை இயல்பாக்கம் செய்யவும் மாற்றவும் இயலும். கூடுதல் முறையாக *Cj* நெறிமுறையும் அதன் மதிப்பெண்கள்

$$\{score(C_j , S_i)\}|_{i=1}^{|SensesD(w)|}$$

தரப்படுகையில் நாம் சொல் w-வின் *i-ஆவது அர்த்தத்தின் நிகழ்தகவு*

$$PCj\,(Si) = \frac{score(C_j, Si)}{\max_k score(C_j, S_k)}$$

என்பதைப் பெறவியலும். இந்த நிகழ்வு தகமைகள் கூட்டப்பட்டு, உயர்ந்த மொத்த மதிப்பெண் பெற்ற அர்த்தம் தெரிந்தெடுக்கப்படுகின்றது:

$$S = \text{argmax}_{Si \in SensesD(w)} \sum_{j=1}^{m} PCj\,(Si).$$

4.2.7.3 தர அடிப்படையிலான ஒருங்கிணைப்பு

முதல் நிரல் வகைப்படுத்தி w என்ற இலக்குச் சொல்லிலுக்கு அர்த்தங்களின் தரத்தைத் தருவதாக எடுத்துக்கொண்டால் தர அடிப்படையிலான ஒருங்கிணைப்புகள் *C*1, *C*2, . . . , *Cm* ஒருங்குமுறைகளால் அதன் தரங்களின் வெளியீட்டின் கூட்டுத்தொகையை மிகைப்படுத்தும் *w*-இன் அர்த்தம் Ŝ-ஐ தெரிந்தெடுப்பதைக் கொண்டிருக்கின்றது (நாம் மிகநல்ல தரத்தைக்கொண்ட அர்த்தம் மிக உயர்ந்த பங்களிப்பைத் தரும்படிக்குத் தரங்களை எதிர்மறை செய்கின்றோம்):

$$S = \text{argmax}_{Si \in SensesD(w)} \sum_{j=1}^{m} -RankC_j\,(S_i),$$

இதில் *RankCj* (*Si*) என்பது வகைப்படுத்தி Cjஆல் *Si*-இன் தரமாகும்.

4.2.7.4 அடாபூஸ்ட் (AdaBoost)

அடாபூஸ்ட் அல்லது மாற்றியமைக்கும் ஊக்கி (adaptative boosting) [Freund and Schapire 1999] பல "வலுவற்ற" வகைப்படுத்திகளின் நேர்கோடான/லீனியார் ஒருங்கிணைப்பாக ஒரு "வலுவான" வகைப்படுத்தியை உருவாக்குவதற்குப் பொதுவான நெறிமுறையாகும்.

4.2.8 மிகக்குறைவான மற்றும் பகுதிகண்காணிக்கப்பட்ட பொருண்மை மயக்கநீக்கம் (Minimally and Semisupervised Disambiguation)

கண்காணிக்கப்பட்ட மற்றும் கண்காணிக்கப்படாத பொருண்மை மயக்க நீக்கம் என்பதற்கு இடையிலான எல்லை தெளிவானதல்ல. நாம் மிகக்குறைவாக அல்லது பகுதி கண்காணிக்கப்பட்ட நெறிமுறைகளை வரையறை விளக்கம் செய்யலாம்; அவை முறையே பொருண்மை வகைப்படுத்திகளை மிகக்குறைவான அல்லது பகுதி மனித கண்காணிப்பு கொண்ட அடையாளப்படுத்தப்பட்ட தரவிலிருந்து கற்கின்றது. சிறிய எண்ணிகையிலான கைகளால் அடையாளப்படுத்தப்பட்ட எடுத்துக்காட்டுகளிலிருந்தான ஒரு தரவுத்தொகுதியின் தானியக்க மேம்படுத்தம் (automatic bootstrapping) மற்றும் ஒருபொருண்மை தொடர்புகளின் (monosemous relatives) பயன்பாடின் அடிப்படையில் இப்பகுதியில் இவ்வகையின் இரண்டு அணுகுமுறைகள் விளக்கப்படுகின்றன

4.2.8.1 மேம்படுத்தல் (bootstrapping)

மேம்படுத்தலின் குறிக்கோள் குறைந்த அளவு பயிற்சித் தரவுடன் ஒரு அர்த்த வகைப்படுத்தியை உருவாக்குவதாகும்; இவ்வாறு கண்காணிப்பின் முக்கியச் சிக்கல்களை நேரிடுவதாகும்: அடையாளப்படுத்தப்பட்ட தரவின் குறைவு மற்றும் தரவு அரிது சிக்கல். மேம்படுத்தல் சில அடையாளப்படுத்தப்பட்ட தரவு A, அடையாளப்படுத்தப்படாத தரவு U-வின் ஒரு பெரிய தரவுத்தொகுதி, ஒன்றோ அதற்குமேலோ அடிப்படை வகைப்படுத்திகள் என்பனவற்றிலிருந்து தொடங்குகின்றது. மேம்படுத்தல் வழிமுறைவரைவின் மீண்டும் மீண்டும் செய்யும் பயன்பாடுகளின் விளைவாக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி A படிப்படியாக வளர்கின்றது/அதிகரிக்கின்றது; அடையாளப்படுத்தப்படாத தரவுக் குழுமம் U-இல் மீதியிருக்கும் எடுத்துக்காட்டுகளுக்கு ஒரு எல்லையை அடைவதுவரை சுருங்குகின்றது. A-இல் உள்ள தொடக்க எடுத்துக்காட்டுகளின் குழுமத்தைக் கைப் புலக்குறிப்பு செய்வதால் [Hearst 1991] அல்லது துல்லிய பொது அறிவின் உதவியுடன் தானியக்கத் தேர்வால் [Yarowsky 1995] உருவாக்க இயலும்.

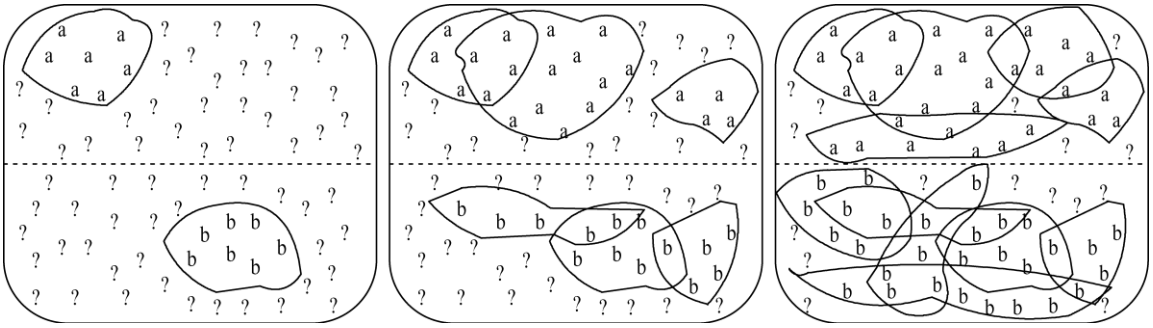
சொற்பொருண்மை மயக்கநீக்கத்தில் மேம்படுத்தலுக்கு இரு முக்கிய அணுகுமுறைகள் உள்ளன: துணைப் பயிற்சி (cotraining), தற்பயிற்சி (self-training).

யரோவ்ஸ்கியின் (Yarowsky 1995) மேம்படுத்தும் நெறிமுறையும் ஒரு தற்பயிற்சி அணுகுமுறையாகும். இது இரண்டு பொது அறிவைச் சார்ந்திருக்கின்றது:

ஒரு சேர்ந்துவருகைக்கு ஒரு அர்த்தம் (Yarowsky 1993): ஒரு சொல்லின் அர்த்தத்தை நிறுவுவதற்கு அடுத்துவரும் சொற்கள் வலுவாகவும் நிரந்தரமாகவும் பங்களிப்பு செய்கின்றது.

ஒரு கருத்தாடலுக்கு ஒரு அர்த்தம் (Gale et al. 1992c): ஒரு சொல் எந்த தரப்பட்ட கருத்தாடலுக்குள்ளும் அல்லது ஆணவத்திற்குள்ளும் நிரந்தரமாக/தொடர்ந்து ஒரே அர்த்தத்தில் குறிப்பிடப்படுகின்றது.

படம் 14: யரோவ்ஸ்கி வழிமுறைவரைவின் எடுத்துக்காட்டு. ஒவ்வொரு மீள் நிகழ்வுக்கும், புதிய எடுத்துக்காட்டுகள் வகுப்பு a-ஆலோ b-ஆலோ புலக்குறிப்பு செய்யப்பட்டுள்ளது மற்றும் பொருண்மை அடையாளப்படுத்தப்பட்ட எடுத்துக்காடுகளின் குழுமம் Aயுடன் சேர்க்கப்படுகின்றது.



நாம் படம் 14இல் யரோவ்ஸ்கியின் வழிமுறைவரைவின் மூன்று மீள்நிகழ்வுகளைக் கூறுகின்றோம்: தொடக்கத்தில் நாம் w என்ற சொல்லுக்கு விதை எடுத்துக்காட்டுகளின் ஒரு சிறிய குழுமம் Aஐ அதன் இரு அர்த்தங்கள் a, b-யுடன் தெரிவுசெய்கின்றோம். தெடர்ந்த மீள்நிகழ்வுகளில் A-ஆல் பயிற்சிசெய்யப்பட்ட தீர்மானப் பட்டியலால் அர்த்தம்-புலக்குறிப்பு செய்யப்பட்ட புதிய எடுத்துக்காட்டுகள் குழுமம் A-யுடன் சேர்க்கப்படுகின்றது. A-யுடன் புதிய புதிய எடுத்துக்காட்டுகள் சேர்க்க இயலாது போகும் போது நாம் நிறுத்துகின்றோம்.

4.3 கண்காணிக்கப்படாத பொருண்மைமயக்கநீக்கம்

கண்காணிக்கப்பட்டாத நெறிமுறைகள் அறிவுப் பேறு முட்டுக்கட்டைகளை விஞ்சுவதற்கு ஆற்றல் கொண்டிருக்கின்றது (Gale et al. 1992b); அதாவது சொற்பொருண்மைக்காகக் கையால் அடையாளப்படுத்தப்பட்ட பெரிய அளவிலான மூலவளங்களின் குறைவை விஞ்சுவதற்கு ஆற்றல் கொண்டிருக்கின்றது. சொற்பொருண்மை மயக்கநீக்கத்திற்கான இவ்வணுகுமுறைகள் ஒரு

சொல்லின் ஒரே பொருண்மை ஒற்றுமையுள்ள அண்மைச் சொற்களைக் கொண்டிருக்கும் என்ற கருத்தை அடிப்படையாகக்கொண்டதாகும். அவைகள் சொல் நேர்வுகளைக் குழுமி உள்ளீட்டுப் பனுவலிலிருந்து சொற் பொருண்மைகளைத் தூண்ட இயலும்; பின்னர் புதிய நேர்வுகளை ஊக்குவிக்கப்பட்ட கொத்துக்களாக (iduced clusters) வகைப்படுத்த இயலும். அவை புலக்குறிப்பு செய்யப்பட்ட பயிற்சிப் பனுவளைச் சார்ந்திருப்பதில்லை; மேலும் தமது சுத்தமான மாதிரியில் அகராதிகள், சொற்களஞ்சியங்கள், மூலப்பொருண்மையியல் ஆய்வுகள் போன்ற இயந்திரம் படிக்கவியலும் மூலவளங்களைப் பயன்படுத்துவதில்லை. இருப்பினும் முழுவதுமான கண்காணிக்கபடாத ஒழுங்குமுறையின் குறைபாடு, அவை எந்த அகராதியையும் பயன்படுத்தாதன் காரணமாக அவை பொருண்மைகளின் பங்கிடப்பட்ட குறிப்புடை தெரிவடைகளைச் (shared reference inventories) சாந்திருக்க இயலாது.

சொற்பொருண்மை மயக்கநீக்கம் எடுத்துக்காட்டான/தனிச்சிறப்பான ஒரு பொருண்மைப் புலக்குறிப்பு செய்யும் செயலாக, அதாவது ஒரு இலக்குச் சொல்லுக்கு ஒரு பொருண்மைப் புலக்குறிப்பின் வெளிப்படையான நிர்ணயமாக/ஒப்படைப்பாக அடையாளம் காணப்படுவதால், கண்காணிக்கப்படாத சொற்பொருண்மை மயக்கநீக்கம் ஏதாவது இரு நேர்வுகளை அவை ஒரே பொருண்மையைச் சார்ந்தவையா இல்லையா என்பதைத் தீர்மானித்துச் சொற்பொருண்மை வேறுபடுத்தலை, அதாவது ஒரு சொல்லின் நேர்வுகளைப் பல வகுப்புகளாகாகப் பகுப்பதை நோக்கமாகக் கொண்டுள்ளது (Schutze 1998). இதன் விளைவாக இந்நெறிமுறைகள் ஒரு அகராதிப் பொருண்மைத் தெரிவடைவில் (dictionary sense inventory) உள்ள மரபுப் பொருண்மைகளுக்குச் (traditional senses) சமமான கொத்துக்களைக் கண்டுபிடிக்காது. இதன் காரணமாக அவற்றின் மதிப்பீடு பெரும்பான்மையும் மிக்கக்கடினமாகும்: ஒரு பொருண்மைக் கொத்தின் தரத்தை மதிப்பிட நாம் மனிதர்களை ஒவ்வொரு கொத்தின் உறுப்பினர்களையும் நோக்குவதற்கு கேட்கவேண்டும்; மேலும் அவைகள் பங்கிட்டுக்கொள்ளும் உறவின் இயல்பை நிர்ணயிக்கவேண்டும் (எ.கா. வினா ஏடு வழியாக) அல்லது இறுதியிலிருந்து-இறுதிப் பயன்பாடுகளில் (end-to-end applications) கொத்துக்களைப் பயன்படுத்தவேண்டும்; இவ்வாறு பிந்தியதன் நிறைவேற்றத்தின் அடிப்படையில் முந்தியதன் சிறப்புப்பண்பு அளவிடப்படும்.

கண்காணிக்கப்படாத சொற்பொருண்மை மயக்கநீக்க அணுகுமுறைகள் கண்காணிக்கப்பட்ட மற்றும் அறிவு அடிப்படையிலான நெறிமுறைகளிலிருந்து வேறுபட

நோக்கத்தைக் கொண்டிருக்கின்றன; அதாவது பொருண்மைப் புலக்குறிப்புகளை நிர்ணயிப்பதற்குப் பதிலாகப் பொருண்மை கொத்துக்களை அடையாளம் காண்கின்றன. இருப்பினும் பொருண்மை வேறுபடுத்தல் மற்றும் பொருண்மைப் புலக்குறிப்புசெய்தல் என்ற இரண்டும் சொற்பொருண்மை மயக்கநீக்கச் செயலின் துணைச்சிக்கல்களாகும் (Schutze 1998); மற்றும் உருவாக்கப்பட்ட கொத்துக்களைப் பின்னர் வரும் நிலையில் சொல் நேர்வுகளைப் பொருண்மை அடையாளப்படுத்தலுக்குப் பயன்படுத்தவியலும் என்ற அளவில் அவைகள் ஒன்றுக்கொன்று கண்டிபான உறவுள்ளவை.

இங்குச் சொற்பொருண்மை மயக்கநீக்கத்தின் முக்கியமான அணுகுமுறைகள் விளக்கப்படும்; அவையாவன: சூழல் கொத்தாக்கம் அடிப்படையிலான நெறிமுறைகள் (methods based on context clustering), சொல் கொத்தாக்கம் (word clustering) மற்றும் உடன்வருகை வரைபடங்கள் (co-occurrence graphs).

4.3.1 சூழல் கொத்தாக்கம்

கண்காணிக்கப்படாத அணுகுமுறையின் முதல் குழுமம் சூழல் கொத்தாக்கத்தின் சாயல் அடிப்படையிலனதாகும். ஒரு தரவுத்தொகுதியில் ஒரு இலக்குச் சொல்லின் ஒவ்வொரு நேர்வும் ஒரு சூழல் வெக்டார் (context vector) ஆக உருப்படுத்தம் செய்யப்படுகின்றது. வெக்டார்கள் பின்னர் குழுமங்களாகக் கொத்தாக்கம் செய்யப்படுகின்றன; அவை ஒவ்வொன்றும் இலக்குச் சொல்லின் ஒரு பொருண்மையை அடையாளம் காண்கின்றன.

இவ்வகையின் ஒரு வரலாற்று அணுகுமுறை சொல் இடப்பரப்பு (word space) அடிப்படையிலானதாகும் (Schütze 1992); அதாவது சொற்கள் பரிமாணங்களாக வருகிற வெக்டார் இடப்பரப்பு (vector space). ஒரு தரவுத்தொகுதியில் உள்ள w என்ற ஒரு சொல்லை ஒரு வெக்டார் ஆக உருப்படுத்தம் செய்யலாம்; இதன் j-ஆவது கூறு ஒரு நிலையான சூழலுக்குள் (ஒரு வாக்கியம் அல்லது ஒரு பெரிய சூழல்) w என்ற சொல்லுடன் wj என்ற சொல் எத்தனை தடவை சேர்ந்து வருகின்றது என்பதைக் கணக்கிடுகின்றது. இம்மாதிரியின் உள்ளுறையும் கருதுகோள் என்னவென்றால் சொற்களின் வினியோகம் பற்றிய சுருக்கக்குறிப்பு உட்படையாக அவற்றின் பொருண்மையியலை வெளிப்படுத்துகின்றது.

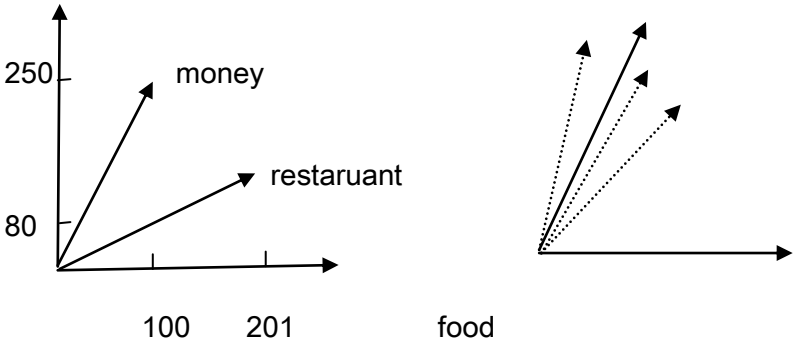
15அ என்ற படம் restaurant = (210, 80) மற்றும் money = (100, 250) என்ற சொல் வெக்டார்களின் இரு எடுத்துக்காட்டுகளைக் காட்டுகின்றது; இதில் முதல் பரிமாணம் food என்ற

சொல்லுடனான சேர்ந்துவருகைகளின் எண்ணிக்கையை உருப்படுத்தம் செய்கின்றது. இரண்டாவது bank என்ற சொல்லுடன்னான சேர்ந்துவருகைகளைக் கணக்கிடுகின்றது.

v மற்றும் *w* என்ற சொற்களுக்கு இடையிலான ஒற்றுமையை வடிவயியல் அடிப்படையில் அளக்க இயலும்; எடுத்துக்காட்டாக *v* மற்றும் *w* பொருத்தமுள்ள வெக்டார்களுக்கு இடையிலுள்ள சொற்களால் அளக்க இயலும்.

படம் 15அ: இரண்டு சொல் வெக்டார்கள் *restaurant* = (210, 80) and *money* = (100, 250). என்பதன் எடுத்துக்காட்டு. 15ஆ. Stock என்பதற்கு ஒரு சூழல் வெக்டார்; இது ஒரே சூழலில் வரும் சொற்களின் வெக்டார்களின் கூட்டுத்தொகை (centroid or the sum) ஆகக் கணக்கிடப்படுகின்றது

bank



4.3.2 சொல் கொத்தாக்கம் (word clustering)

முந்தைய பகுதியில் சொற் பொருண்மைகள் முதலாவது அல்லது இரண்டாவது நிரல் சூழல் வெக்டார்களாக உருப்படுத்தம் செய்யப்பட்டது. சொற் பொருண்மைகளின் தூண்டும் ஒரு வேறுபட்ட அணுகுமுறை சொல் கொத்தாக்க நுட்பங்களை (word clustering techniques) கொண்டது; அதாவது பொருண்மையியல் அடிப்படையில் ஒற்றுமையுள்ள சொற்களைக் கொத்தாக்கம் செய்வதை நோக்கமாகக்கொண்ட நெறிமுறைகள் கொண்டது; இவ்வாறு ஒரு குறிப்பிட்ட அர்த்தத்தை வெளிப்படுத்தும்.

சொல் கொத்தாக்கத்தின் ஒரு மிக அறியப்பட்ட அணுகுமுறை w_0 என்ற ஒரு இலக்குச் சொல்லுக்கு $W = (w_1..., w_k)$ என்ற ஒற்றுமையுள்ள சொற்களை (சாத்தியமாக ஒருபொருள் பன்மொழியச் சொற்களை) அடையாளம் காணுதலை உள்ளடக்கும். w_0 என்ற சொல்லுக்கும் w_i என்ற சொல்லுக்கும் இடையிலுள்ள ஒற்றுமை ஒரு தரவுத்தொகுதியில் நேரும் தொடரியல் சார்பால் தரப்பட்ட அவற்றின் தனிநிலைப் பண்புக்கூறுகளின் (எ.கா. எழுவாய்-வினை, வினை-

செயப்படுபொருள், பெயரடை-பெயர் போன்றன) தகவல் பொருளடக்கம் அடிப்படையில் நிர்ணயிக்கப்படுகின்றது. இவ்விரு சொற்கள் பங்கிட்டுக்கொள்ளும் சார்புகள் கூடும்போது தகவல் பொருடக்கமும் மேம்படும். இருப்பினும் சூழல் வெக்டார்களைப் பொறுத்தவரையில் W என்பதில் உள்ள சொற்கள் w_0 என்பதன் எல்லாப் பொருண்மைகளையும் உள்ளடக்கும். பொருண்மைகளை வேறுபடுத்த ஒரு சொல் கொத்தாக்க வழிமுறைவரைவு (word clustering algorithm) பயன்படுத்தப்படுகின்றது. W என்பது w_0 என்பதுடன் உள்ள ஒற்றுமையின் அளபால் நிரல்படுத்தப்பட்ட ஒற்றுமையான சொற்களின் பட்டியலாக இருக்கட்டும். T என்ற ஒற்றுமைக் கிளை (similarity tree) தொடக்கத்தில் உருவாக்கப்படுகின்றது; இது w_0 என்ற ஒரு தனிநிலை கணுவைக் கொண்டிருக்கும். அடுத்தபடியாக ஒவ்வொரு $i \in \{1,\ldots,k\}$ என்பதற்கும், $w_i \in W$ என்பது T கிளையில் wj-இன் ஒரு குழந்தையாக சேர்க்கப்படுகின்றது; இதன் படி $\{w_0, \ldots w_{i-1}\}$ என்பனவற்றில் w_j என்பது w_i என்பதற்கு மிக ஒற்றுமையுள்ள சொல்லாகும். ஒரு மிகை அகற்றும் நடைமுறைக்குப் பின் w_0 என்பதில் வேர்விடும் ஒவ்வொரு துணைக் கிளையும் w_0 என்பதன் ஒரு தனிப்பட்ட (வேறுபட்ட) பொருண்மையாகக் கருதப்படுகின்றது.

குழுவினால் செய்யப்படும் கொத்தாக்கம் (clustering by committee (CBC) வழிமுறைவரைவு (algorithm) [Lin and Pantel 2002] என்று அழைக்கப்படுகின்ற ஒரு பின்தொடரும் அணுகுமுறையில் மற்றொரு சொல் கொத்தாக்க நெறிமுறை (different word clustering) முன்மொழியப்பட்டது. திரும்பவும் ஒற்றுமையைக் கணக்கிட ஒவ்வொரு சொல்லும் ஒரு பண்புக்கூறு வெக்டராக உருப்படுத்தம் செய்யப்படுகிறது; இதில் ஒவ்வொரு பண்புக்கூறும் அச்சொல் நேரும் தொடரியல் சூழலின் வெளிப்படுத்தமாகும். இலக்குச் சொற்களின் ஓர் குழுமம் தரப்படுகையில், Sij என்பது wi மற்றும் wj என்ற சொற்களுக்கிடையில் இணைவயமான ஒற்றுமையைக் கொண்டிருக்கின்றது எனும்படி S என்ற ஒரு ஒற்றுமைச் சட்டகம் உருவாக்கப்படுகின்றது.

E என்ற சொற்களின் ஒரு குழுமம் தரப்படுகையில் இரண்டாவது நடைப்பாங்காக E-இல் குழுக்கள் (committees) என்று அழைக்கப்படுகின்ற சொற்களின் கொத்துக்களின் குழுமங்களை நிர்ணயிக்க ஒரு மறுதரவு/மறுநிகழ்வு செயல்பாடு பயன்படுத்தப்படுகின்றது. இந்த இறுதிக்கு ஒரு நிலைபேறு பெற்ற கொத்தாக்க நுட்பத்தை அதாவது சராசரி-இணைப்பு கொத்தாக்கம் (average-link) பயன்படுத்தப்படுகின்றது. ஒவ்வொரு நடைப்பாங்கிலும் எந்தக் குழுவாலும்

உட்படுத்தப்படாத மீதிச் சொற்கள் [அதாவது ஒவ்வொரு குழுவின் சென்ட்ராய்டுடன் (centroid) ஒற்றுமை இல்லாத சொற்கள்] அடையாளம் காணப்படுகின்றன. மீதிச் சொற்களிலிருந்து கூடுதலான குழுக்களைக் கண்டுபிடிக்க மறுதரவு/மீள்நிகழ்வு முயற்சிகள் எடுக்கப்படுகின்றன. முன்னர் உள்ளதுபோல் ஒவ்வொரு சொல்லும் ஒரு தனிநிலைக் குழுவைச் சார்வதால் குழுக்கள் பொருண்மைகளைச் சுருக்குகின்றன (conflates).

இறுதியாக ஒரு பொருண்மை வேறுபடுத்தும் நடைப்பாங்காக, ஒரு பண்புக்கூறு வெக்டாராக மீண்டும் உருப்படுத்தம் செய்யப்பட்டுள்ள ஒவ்வொரு இலக்குச் சொல் w Є E யும் ஒவ்வொரு குழுவின் சென்ட்ராய்டுடன் அதற்கு இருக்கும் ஒற்றுமை அடிப்படையில் மீண்டும் மீண்டும் அதன் கூடுதல் ஒற்றுமையுள்ள கொத்திற்கு ஒதுக்கப்டும். w என்ற ஒரு சொல் c என்ற ஒரு குழுவிற்கு ஒதுக்கப்பட்ட பின்னர், w-க்கும் c-யில் உள்ள தனிமங்களுக்கும் இடையிலுள்ள மேலுறும் பண்புக்கூறுகள் w-இன் உருப்படுத்தத்திலிருந்து நீக்கப்படுகின்றன; இது பின்னர் வரும் மீள்நிகழ்வில் அதே சொல்லின் பெரும்பான்மை குறைந்த பொருண்மைகளை அடையாளம் காண்பதை அனுமதிக்கின்றது.

61% துல்லியம் மற்றும் 51% மீளழைப்பைப் பெற்றுச் சொல்வலை சொல் பொருண்மைகளை (wordNet word senses) அடையாளங்காணும் செயலின் மீது CBC மதிப்பீடு செய்யப்படுகின்றது. முந்தைய அணுகுமுறைகளுக்கு முரணாக, CBC கருத்துருக்களின் தட்டையான பட்டியலை வெளியிடுகின்றது (அதாவது அது கொத்துக்களுக்கு ஒரு படிநிலை அமைப்பைத் தருவதில்லை). சமீபத்தில் சொல் மும்மடங்குகள் (word triplets) அடிப்படையில் சொற்பொருண்மை தூண்டலை நிறைவேற்றும் ஒரு புதிய அணுகுமுறை அறிமுகப்படுத்தப்பட்டுள்ளது (Bordag 2006). "ஒரு சேர்ந்துவருகைக்கு ஒரு பொருண்மை" என்ற ஊகத்தையும் பண்புக்கூறுகளாகத் தங்கள் மேலுறல்களைப் (intersections) பயன்படுத்தும் கொத்துக்களின் சேர்ந்துவருகை மும்மடங்குகளையும் இந்நெறிமுறை சார்ந்திருக்கின்றது. பொருண்மை தூண்டல் உயர்ந்த துல்லியத்துடன் நிறைவேற்றப்படுகின்றது (மீளழைப்பு சொல்வகைப்பாடையும் நிகழ்வெணையும் பொறுத்து வேறுபடுகின்றது).

4.3.3 சேர்ந்துவருகை வரைபடங்கள் (co-occurrence graphs)

சொற் பொருண்மை வேறுபடுத்தலின் ஒரு வேறுபட்ட பார்வை வரைபட-அடிப்படை அணுகுமுறைகளால் (graph-based approaches) தரப்படுகின்றது. இவ்வணுகுமுறைகள் ஒரு

சேர்ந்துவருகை வரைபடக் (co-occurrence graph) கருத்துச்சாயல் அடிப்படையிலானது; அதாவது V என்ற வெர்ட்டிஸ்சுகள் ஒரு பனுவலில் உள்ள சொற்களுடன் பொருந்துகின்ற மற்றும் E என்ற விளிம்புகள் ஒரே பத்தியில் அல்லது ஒரு பெரிய சூழலில் ஒரு தொடரியல் உறவில் சேர்ந்துவரும் சொற்களின் இணைகளை இணைக்கின்ற G = (V, E) என்ற ஒரு வரைபடம் அடிப்படியிலானது.

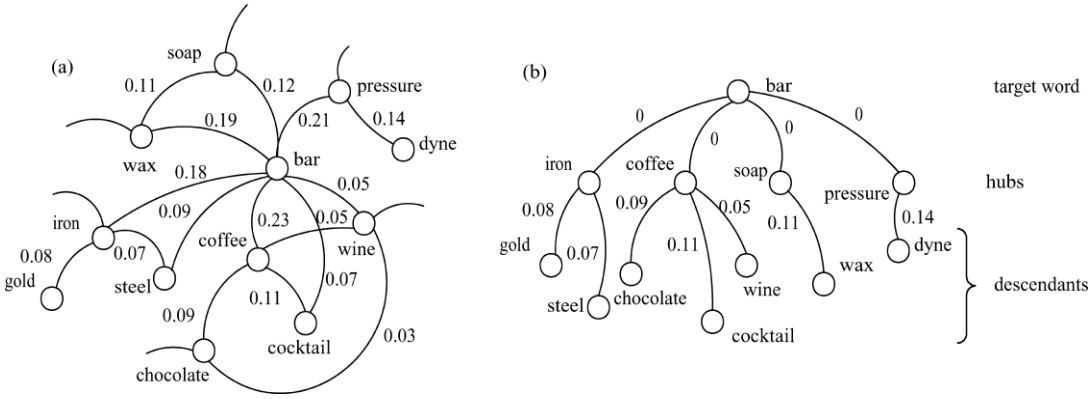
சூழலில் சொற்களுக்கு இடையில் இலக்கண உறவுகள் அடிப்படையிலான ஒரு சேர்ந்துவருகை வரைபடத்தின் (cooccurence graph) உருவாக்கம் விட்டோஸ் மற்றும் டொரோ (Widdows and Dorow 2002) மற்றும் விட்டோஸ் (Widdows 2003) என்பவர்களால் விளக்கப்பட்டுள்ளது. w என்ற பொருண்மை மயக்கமுள்ள ஒரு இலக்குச் சொல் தரப்படுகையில் w-ஐச் சுற்றி G_w என்ற ஒரு வட்டார வரைபடம் உருவாக்கப்படுகின்றது. G_w என்பதுடன் தொடர்புடைய அண்மையிருப்பு வார்ப்பை (adjacency matrix) இயல்பாக்கத்தால் மார்க்கோவ் சங்கலியாக (Markov chain) அந்த வரைபடத்தை நாம் பொருள்கோள் செய்யலாம். கூடுதலான புதிய தூர அண்மைபட்டவைகளை ஆய்வதையும் கூடுதல் புகழ்பெற்ற கணுக்களுக்கு ஆதரவு தருவதையும் நோக்கமாகக் கொண்ட ஒரு விரிவடையும் மற்றும் ஒரு அதிகரிக்கும் நடைபாங்கு அடிப்படியில் மார்கோவ் கொத்தாக்க வழிமுறைவரைவு (van Dongen 2000) சொல் பொருண்மைகளை நிர்ணயிக்கப் பயன்படுத்தப்படுகின்றது.

இதைத் தொடர்ந்து வெரொனிஸ் (Véronis) ஒரு ஹைபர்லெக்ஸ் (HyperLex) என்ற ஒரு குறிப்பிட்ட பணிக்குரிய அணுகுமுறையை முன்மொழிந்தர். முதலில் இலக்குச் சொல் வரும்/நேரும் ஒரு பனுவலின் தரவுத்தொகுதியின் (text corpus) பத்திகளில் நேரும்/வரும் சொற்கள் தாம் கணுக்கள் எனும்படியும் அவ்வரைபடத்தில் அவை ஒரே பத்தியில் சேர்ந்துவருகை செய்தால் சேர்க்கப்படும் சொற்களின் ஒரு இணைகளுக்கு இடையில் ஒரு விளிம்பும் ஒரு சேர்ந்துவருகை வரைபடம் உருவாக்கப்படுகின்றது. ஒவ்வொரு விளிம்பும் விளிம்பால் இணைக்கப்பட்ட இரு சொற்களின் ஒத்தறி சேர்ந்துவருகை நிகழ்வெண் அடிப்படையில் ஒரு வெயிட் ஒத்துக்கப்படுகின்றது.

இரண்டாவது நடைப்பாங்காக ஒரு மீள்நிகழும் வழிமுறை வரைவு சேர்ந்துவருகை வரைபடத்திற்கு பயன்படுத்தப்படுகின்றது: ஒவ்வொரு மீள்நிகழ்விலும் வரைபடத்தில் மிக உயர்ந்த ஒத்தறி அளபு கொண்ட கணு ஒரு ஹஃப் (hub) ஆகத் (மூலப் பனுவலில் ஒரு கணுவின் அளபும் அதன் நிகழ்வெண்ணும் மிக உயர்ந்த அளவில் பொருத்தப்படுகின்றது என்ற

பரிசோதனை கண்டுபிடிப்பு அடிப்படையில்) தெரிந்தெடுக்கப்படுகின்றது. இதன் விளைவாக அதன் எல்லா அண்மைப்பட்டவைகளும் ஹஃப் அங்கத்தினர்களாகத் தகுதி பெறவில்லை. தெரிந்தெடுக்கப்பட்ட ஹஃப்களின் முழுக் குழுமமும் ஆர்வமுள்ள சொல்லின் பொருண்மைகளை உருப்படுத்தம் செய்வதாகக் கூறப்படுகின்றது. ஹஃப்கள் இலக்குச் சொல்லுடன் பூஜிய-எடை விளிம்புகளால் (zero-weight edges) தொடர்புபடுத்தப்பட்டுள்ளது மற்றும் முழு வரைபடத்தின் குறைந்த இட அளவுக் கிளை (minimum spanning tree (MST) கணக்கிடப்படுகின்றது (எ.கா. படம் 16(b)).

படம் 16அ சேர்ந்துவருகை வரைபட த்தின் பகுதி. படம் 16ஆ: bar$_n$ இலக்குச் சொல்லுக்கு குறைந்த இட அளவுக்கிளை



4.4. அறிவு-அடிப்படையிலான பொருண்மை மயக்கநீக்கம் (Knowledge-based disambiguation)

அறிவு-அடிப்படையிலான அல்லது அகராதி அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் சூழல்களில் சொற்களின் பொருண்மைகளை ஊகிக்க அறிவு மூலவளங்களைப் (அகராதிகள், சொற்களஞ்சியங்கள், மூலப்பொருண்மையியல் ஆய்வுகள், உடன்வருகைகள், போன்றவை) பயன்படுத்துவதாகும். இந்நெறிமுறைகள் பெரும்பான்மையும் அவற்றின் கண்காணிக்கப்பட்ட மாற்றுக்களைவிட (supervised alternative) குறைந்த அளவு நிறைவேற்றத்தைக் கொண்டிருக்கின்றன; ஆனால் அவை மிகப் பெரிய அளவிலான அறிவு மூலங்களைப் பயன்படுத்துவதால் அவற்றிற்கு அகன்ற/விரிந்த பயன்பாட்டுப் பரப்பு இருக்கின்றது.

சொற்பொருண்மை மயக்கநீக்கத்திற்கு முதல் அறிவு-அடிப்படையிலான அணுகுமுறைகளின் பயன்பாடு 1970-1980களின் காலகட்டத்திற்குக் கொண்டுசெல்லும்; அப்போது சோதனைகள் மிக எல்லைக்குட்பட்ட பொருட்புலங்களில் தான் செயல்முறை

படுத்தப்பட்டது. இச்செயல்களின் அளவீடு அச்சமயத்தில் முதன்மையான கடினமாகும்: பெரும் அளவிலான கணினிசார் மூலவளங்களின் குறைவானது சரியான மதிப்பீடு, ஒப்பீடு மற்றும் இறுதியிலிருந்து-இறுதி பயன்பாடுகளில் இந்நெறிமுறைகளின் பயன்பாடு என்பனவற்றைத் தடைசெய்தது.

இப்பகுதியில் நாம் முக்கியமான அறிவு அடிப்படையிலான நுட்பங்களை மேலோட்டமாகப் பார்ப்போம்: பொருண்மை வரையறை விளக்கங்களின் மேலுறல், தேர்வுக்கட்டுப்பாடுகள் மற்றும் அமைப்புசார் அணுகுமுறைகள் (பெருண்மைசார் ஒற்றுமை அளவீடுகள் மற்றும் வரைபட-அடிப்படையிலான நெறிமுறைகள்). பொரும்பாலான அணுகுமுறைகள் சொல்வலை அல்லது பிற மூலவளங்களிலிருந்து கிடைக்கின்ற தகவல்களைப் பயன்படுத்துகின்றன மான்னிங் மற்றும் ஸ்கூட்ஸ் (Manning and Schütze 1999) மற்றும் மிஹால்செயா (Mihalcea 2006). என்பவர்களின் ஆய்வுகளில் அறிவு அடிபடையிலான அணுகுமுறைகளின் மதிப்புரையைக் காணலாம்.

4.4.1 பொருண்மை வரையறைவிளக்கங்களின் மேலுறல் (Overlapping of sense definition)

ஒரு எளிய மற்றும் உள்ளுணர்வுசார் அறிவு அடிப்படையிலான அணுகுமுறைகள் இரண்டு அல்லது அதற்கு மேலான இலக்குச் சொற்களின் பொருண்மை வரையறைவிளக்கங்களுக்கு இடையிலான சொல் மேலுறலின் கணக்கீட்டைச் சார்ந்துள்ளது. இவ்வணுகுமுறை பொருள்விளக்க மேலுறல் (gloss overlap) அல்லது (லெஸ்க் என்ற படைப்பாளியின் (Lesk 1986) பெயரைப் பின்பற்றி லெஸ்க் வழிமுறை வரைவு (Lesk algorithm) எனப் பெயரிட்டு அழைக்கப்படுகின்றது. ஒரு இரு-சொல் சூழல் (two-word context) தரப்படுகையில் (w_1, w_2) மிக உயர்ந்த மேலுறல் உள்ள வரையறை விளக்கங்களைக் கொண்ட இலக்குச் சொற்களின் பொருண்மைகள் உள்ளவை சரியானவைகளாக ஊகிக்கப்படுகின்றன. முறையாக w1, w2 என்ற இரண்டு சொற்கள் தரப்படுகையில், பின்வரும் மதிப்பெண்கள் சொற் பொருண்மைகளின் ஒவ்வொரு இணைகளுக்கும் கணக்கிடப்படுகின்றது S1 Є Senses (w1) and S2 Є Senses (w2):

Score Lest (S1,S2) = |gloss (S1) ∩ gloss(S2)|,

இதில் gloss(Si) என்பது wi-இன் Si என்ற பொருண்மையின் பனுவல் வரையறைவிளக்கம். மேற்கண்ட வாய்ப்பாட்டை அதிகப்படுத்தும் பொருண்மைகள் முறையே சொற்களுக்கு

ஒதுக்கப்படும். இருப்பினும் இது பொருள்விளக்க மேலுறல்களின் கணக்கீட்டை வேண்டும் |Senses(w1)|.|Senses(w2)|.

4.4.2 விருப்பத்தேர்வுகள் (Selectional preferences)

அறிவு அடிப்படையிலான வழிமுறைவரைவுகளின் ஒரு வரலாற்று வகை சூழலில் வரும் ஒரு இலக்குச் சொல்லின் அர்த்தங்களின் எண்ணைக்கையைக் கட்டுப்படுத்த விருப்பத்தேர்வுகளைப் பயன்படுத்துகின்றது. விருப்பத்தேர்வுகள் அல்லது கட்டுப்பாடுகள் வாக்கியங்களில் ஒரு சொல் பொருண்மை அச்சொல் சேரும் சொற்களின் மீது சுமத்தும் பொருண்மை வகையின் கட்டுப்பாடுகள் ஆகும் (பெரும்பான்மையும் இலக்கண உறவுகள் வழியாக ஏற்படுவது). எடுத்துக்காட்டாக eat 'உண்' என்ற வினை ஒரு விலங்கு இருப்புப்பொருளை எழுவாயாகவும் உண்ணும் இருப்புப்பொருளை நேரடி செயப்படுபொருளாகவும் எதிர்பார்க்கின்றது. நாம் தேர்வுக்கட்டுப்பாடுகளுக்கும் விருப்பத்தேர்வுகளுக்கும் இடையில் உள்ள வேறுபாட்டைக் காணலாம்; முந்தையது கட்டுப்பாடுகளை மீறும் பொருண்மைகளைத் தள்ளிவிடுகின்றது; பிந்தையது (தற்கால அனுபவாத செயல்களின் கூடுதல் எடுத்துக்காட்டானது) தேவைகளைத் திருப்திசெய்யும் பொருண்மைகளைத் தேர்வுசெய்யும் போக்குகொண்டது.

விருப்பத்தேர்வுகளைக் கற்கும் மிக எளிதான வழி ஒரு சொல்லுக்குச்சொல் உள்ள உறவால் (word-to-word relation) தரப்படுகின்ற சேர்க்கையின் பொருண்மை பொருத்தத்தை நிர்ணயிப்பதாகும். இவ்வகையின் மிக எளிய அளவீடு நிகழ்வெண் கணக்கீடாகும். w1, w2 என்ற சொற்களின் ஒரு இணையும் R என்ற தொடரியல் உறவும் (எ.கா. எழுவாய்-வினை, வினை-செயப்படுபொருள் போன்றன) தரப்படுகையில், இந்நெறிமுறை ஒரு பகுத்துக்குறிக்கப்பட்ட பனுவலின் தரவுத்தொகுதியில் இந்நேர்வுகளின் எண்ணிகையை (R, w_1, w_2) கணக்கிடுகின்றது (பார்க்க எ.கா. Hindle and Rooth 1993). ஒரு சொலுக்குச்சொல் உறவின் பொருண்மைப் பொருத்தத்தின் மற்றொரு மதிப்பீடு, (estimation) w_2 என்ற சொல்லும் R என்ற உறவும் தரப்படுகையில் w_1 என்ற சொல்லின் கட்டுப்பாட்டுச் சாத்தியம் (conditional probability) ஆகும்:

$$R{:}P(w_1|w_2,R) = \frac{\text{Count}(w_1,w_2,R)}{\text{Count}(w_2,R)}$$

சொல்லுக்குச்சொல் (word-to-word) அல்லது வகுப்புக்குவகுப்பு (class-to-class) மாதிரிகளைத் தருவதற்கு, அதாவது பொருண்மை வகுப்புகளுக்கு அறிவுப் பேறை பொதுமையாக்கம் செய்ய மற்றும் தரவு அரிது சிக்கலை (data spaceness problem) விடுவிக்கச் சொல்வலை போன்ற கையால் செய்யப்பட்ட வகைப்பாட்டியல்களைச் சொற்களிலிருந்து கருத்துரு வகுப்புகளை ஆக்கப் பயன்படுத்தலாம். பல நுட்பங்கள் உருவாக்கப்பட்டுள்ளன: தேர்வு ஒன்றிணைப்பு அளவீடுகள் (measures of selectional association) (Resnik 1993, 1997), குறைந்த விளக்க நீளத்தைப் பயன்படுத்தும் கிளை வெட்டு மாதிரிகள் (tree cut models using the minimum description length) (Li and Abe 1998; McCarthy and Carroll 2003), மறைக்கப்பட்ட மார்கவ் மாதிரிகள் (hidden markov models (HMMs) (Abney and Light 1999), வகுப்பு அடிப்படையிலான சாத்தியம் (class-based probability) (Clark and Weir 2002; Agirre and Martinez 2001), பெய்சன் வலைப்பின்னல்கள் (Bayesian networks) (Ciaramita and Johnson 2000) போன்றன. எல்லா அணுகுமுறைகளும் பெரிய தரவுத்தொகுதிகளைப் பயன்படுத்துகின்றன; மற்றும் பயனிலைகளின் பங்கெடுப்பாளர்களின் பொருண்மை வகுப்புகளைப் பற்றிய அறிவுடன் உற்றுநோக்கப்பட்ட நிகழ்வெண்களை ஒன்றுசேர்த்து அப்பயனிலைகளின் விருப்பத் தேர்வுகளை மாதிரிப்படுத்துகின்றது.

சொல்லுக்குச் சொல், சொல்லுக்கு வகுப்பு, வகுப்புக்கு வகுப்பு அணுகுமுறைகளின் ஒரு ஒப்பீடு அகிர்ரெ மற்றும் மார்டினெஸ் (Agirre and Martinez 2001) என்போரால் முன்வைக்கப்பட்டுகின்றது; அவர்கள் நாம் முந்தைய மாதிரிகளிலிருந்து பிந்தைய மாதிரிகளுக்குச் செல்லுகையில் பயன்பாட்டுப் பரப்பு வளர்கின்றது (சொல்லுச்சொல் விருப்பத்தேர்வுகளுக்கு 26%; சொல்லுக்கு-வகுப்புக்கு 86%, வகுப்புக்கு-வகுப்புக்கு நெறிமுறைகள் 97.3%) மற்றும் இவ்வாறே துல்லியம் குறைகின்றது (முறையே 95%-இலிருந்து 66.9%-க்கு அதிலிருந்து 66.6%-க்கு) என்று காண்கின்றனர்.

பொதுவாகத் தேர்வுக்கட்டுப்பாடு அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்க அணுகுமுறைகள் லெஸ்க் அடிப்படியிலான நெறிமுறைகள் (Lesk-based methods) அல்லது கூடுதல் நிகழ்வெண் பொருண்மை உய்த்தறிவு (most frequent sense heuristic) போன்று நன்றாக நடைபெறுவதாகக் தெரியவில்லை.

4.4.3 அமைப்புமுறை அணுகுமுறைகள் (structural approaches)

சொல்வலை போன்ற கணினிசார் பேரகராதிகள் கிடைப்பதன் காரணமாக அம்மாதிரியான பேரகராதிகளில் கிடைக்கும் கருத்துரு வலைப்பின்னலின் அமைப்பை ஆயவும் பயன்படுத்தவும் பல அமைப்புமுறை அணுகுமுறைகள் உருவாக்கப்பட்டன. வட்டார மற்றும் உலகவய சூழலில் அமைப்பொழுங்குகளின் புரிந்துகொள்ளலையும் அளவீட்டையும் அமைப்புமுறை அமைப்பொழுங்கு புரிந்துகொள்கையின் களத்தில் ஒன்றாகச் சேர்த்துவைக்க இயலும் (Fu 1982; Bunke and Sanfeliu 1990); இது பண்புக்கூறுகளின் அமைப்புமுறை பரஸ்பர உறவுகள் (structural interrelationships) அடிப்படையில் தரவை வகைப்படுத்துவதை நோக்கமாகக் கொண்டது. நாம் இவ்வகையின் இரு முக்கிய அணுகுமுறைகளை முன்வைக்கின்றோம்: ஒற்றுமை அடிப்படையிலான மற்றும் வரைபடம் அடிப்படையிலான நெறிமுறைகள்.

4.4.3.1 ஒற்றுமை அளவீடுகள் (similarity measures)

1990களின் தொடக்கத்தில் சொல்வலை அறிமுகப்படுத்தப்பட்டபோது சொற் பொருண்மைகளுக்கு இடையிலான பொருண்மைத் தொடர்புகளின் வலையமைப்பைப் பயன்படுத்த வேண்டி பொருண்மை ஒற்றுமைகளின் பல அளவீடுகள் உருவாக்கப்பட்டன. பொருண்மை ஒற்றுமையின் அளவீடு தரப்படுகையில், அதைப் பின்வருமாறு வரையறை விளக்கம் செய்யலாம்:

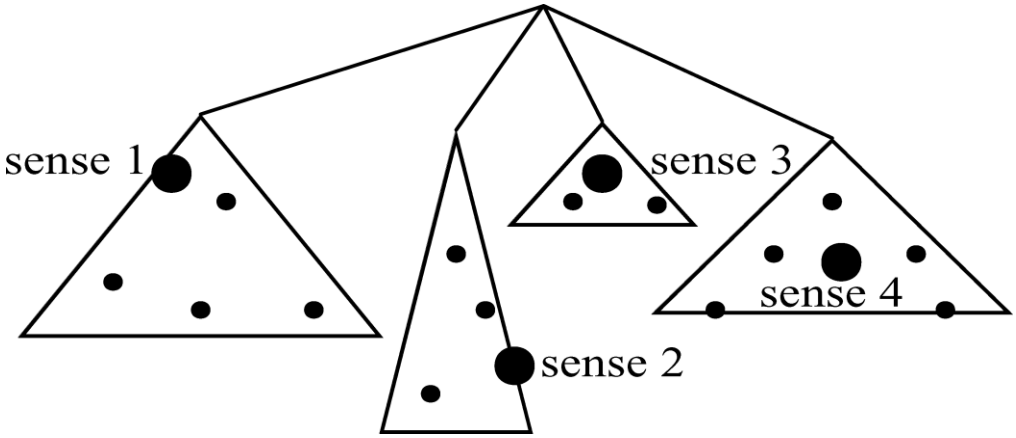
Score:SenseD X Sense D $\rightarrow$ [0, 1]

இதில் SenseD என்பது ஒரு குறிப்புரை பேரகராதியில் பட்டியலிடப்பட்டுள்ள பொருண்மைகளின் முழு குழுமம் ஆகும்; நாம் பொதுவான பொருண்மை மயக்கநீக்கத்தை நமது ஒற்றுமை அளவீடு அடிப்படையில் வரையறை விளக்கம் செய்யலாம். நாம் T= (w1…wn) என்ற பனுவலில் உள்ள w1 என்ற ஒரு இலக்குச் சொல்லை பின்வரும் கூட்டுத்தொகையைப் பெரிதாக்கும் Wi-இன் S என்ற பொருண்மையை விருப்பத்தேர்வு செய்து பொருண்மை மயக்கநீக்கம் செய்யலாம்.

படம் 17-இல் நாம் கருத்துரு அடர்த்தியின் அடிப்படைக் கருத்தைக் காட்டுகின்றோம். நாம் இலக்குச் சொல் w என்பதன் நான்கு அர்த்தங்களைப் பெரிய புள்ளிகளால் சுட்டிக்காட்டுகின்றோம். எடுத்துக்காட்டில் w என்பதன் ஒவ்வொரு அர்த்தமும் சொல்வலை பெயர்ச்சொல் வகைப்பாட்டியலின் வேறுபட்ட துணைப் படிநிலை அமைப்பைச் சாரும்.

படம் 17: ஒரு சொற் சூழலுக்கு கருத்துருசார் அடர்த்தியின் எடுத்துக்காட்டு; இது இலக்குச் சொல்லின் நான்கு அர்த்தங்களை உட்படுத்தும். சூழலில் வரும் சொற்களின் அர்த்தங்கள் சிறிய

புள்ளிகளாலும் இலக்குச் சொல்லின் அர்த்தங்கள் பெரிய புள்ளிகளாலும் உருப்படுத்தம் செய்யப்பட்டுள்ளன.



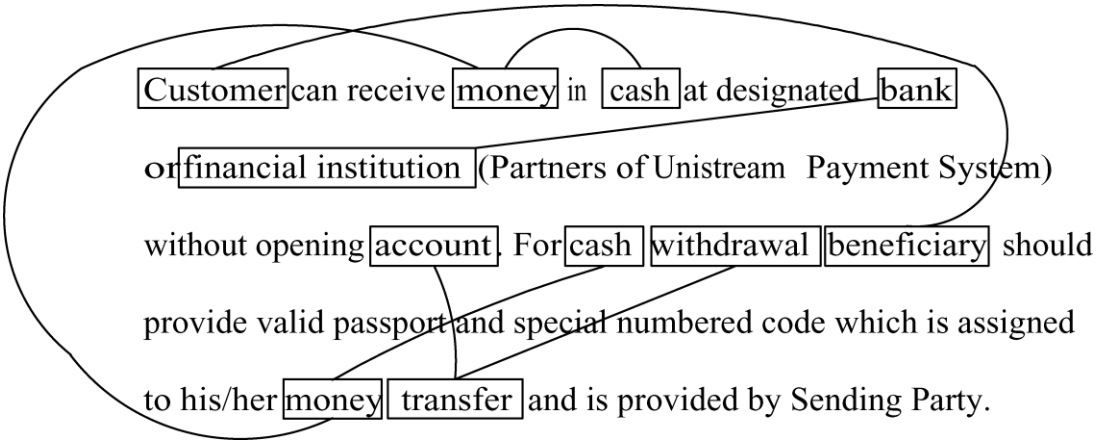
4.4.3.2 வரைபட அடிப்படையிலான அணுகுமுறைகள் (Graph based approaches)

இப்பகுதியில் சூழலில் சொற்களின் மிகப்பொருத்தமான பொருண்மைகளை நிர்ணயிப்பதற்கு வரைபட அமைப்புகளைப் பயன்படுத்துவதன் அடிப்படையில் அமைந்த பல அணுகுமுறைகள் முன்வைக்கப்படுகின்றன. பெரும்பான்மையான இவ்வணுகுமுறைகள் சொற்சங்கலியின் கருத்துச்சாயலுடன் தொடர்புள்ளனவாகவோ அல்லது சொற்சங்கலியின் (lexical chain) கருத்துச்சாயால் ஊக்குவிக்கப்பட்டனவாகவோ இருக்கின்றன. ஒரு சொற்சங்கலி (Halliday and Hasan 1976; Moris and Hirst 1991) என்பது ஒரு பனுவலில் உள்ள பொருண்மையியல் அடிப்படையில் தொடர்புள்ள $w_1 \ldots w_n$ என்ற சொற்களின் கோர்வையாகும்; இதன்படி w_i என்பது w_i+1 என்பதுடன் ஒரு சொல்-பொருண்மை உறவால் தொடர்புள்ளது (எ.கா. is-a, has-part போன்ற உறவுகள்). எடுத்துக்காட்டாக பின்வருவன சொற் சங்கலியின் எடுத்துக்காடுகளாகும்: Rome → city → inhabitant, eat→ dish→ vegetable→ aubergine.

இவ்வமைப்புகள் கருத்தாடல் இயைபின் பகுப்பாய்வு (analysis of discourse cohesion) (Morris and Hirst 1991), பனுவல் சுருக்கமாக்கம் (text summarization) (Barzilay and Elhadad 1997), சொற்குளறுபடிகளின் திருத்தம் (correction of malapropism) (Hirst and St-Onge 1998) போன்றவைகளுக்குப் பயன்படுத்தப்படுகின்றன.

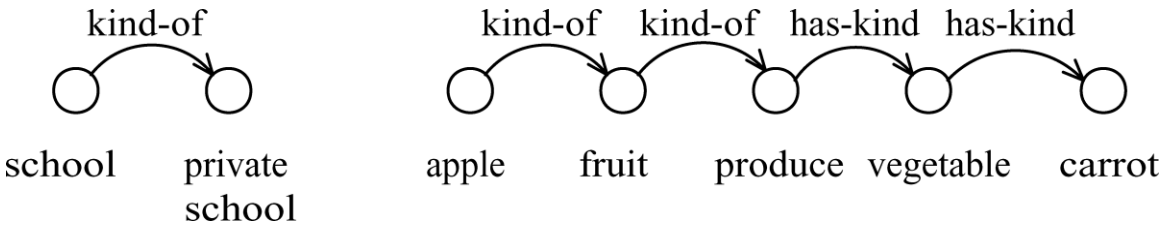
நாம் சொற்சங்கலிகளைப் பொருண்மை ஒற்றுமையின் அளவீடுகளின் ஒரு உலகவய மாற்றமைப்பாகக் காணலாம்; பொருண்மை ஒற்றுமையின் அளவீடுகள் மாறாக வட்டார சூழல்களில் பெரும்பான்மையும் பயன்படுத்தப்படுகின்றது. தொடர்புள்ள சொற்களுக்கு இடையில் ஆற்றலுள்ள சொற்சங்கிலிகளுடன் கூடிய பனுவலின் ஒரு பாகத்தை படம் 18 எடுத்துக்காட்டுகின்றது.

படம் 18: பனுவலின் ஒரு பகுதியில் சில சொற் சங்கிலிகள்



சொற் சங்கலிகளின் கருத்துச்சாயலால் ஊக்கப்படுத்தப்பட்ட அணுகுமுறைகளுக்குள் நாம் ஹரபாகியு மற்றும் பிறரால் (Harabagiu et al 1999) செய்யப்பட்ட அணுகுமுறையைக் குறிப்பிடலாம்; இதில் அகராதிப் பொருள் விளக்கங்களைப் பொருண்மை மயக்கநீக்கம் செய்ய சொல்-பொருண்மை ஊகங்கள் (lexicosemantic heuristics) பயன்படுத்தப்படுகின்றது; ஒவ்வொரு ஊகமும் மொழியின் ஒரு குறிப்பிட்ட இயல்நிகழ்ச்சியை (எ.கா. ஒற்றைப்பொண்மை (monosemy), மொழியியல்சார் இணைமை (linguistic parallelism) போன்றன) கையாளுகின்றது; இவற்றில் சிலவற்றைச் சொற் சங்கிலியின் தனிச்சிறப்பான வகைகளாக உருவகிக்க இயலும்.

படம் 19: சொற் சங்கிலிகளின் இரு எடுத்துக்காட்டுகள்



மிஹல்செயா மற்றும் பிறர் (Mihalcea et al 2004) சொற்களஞ்சிய வலையமைப்பின் அமைப்பை

ஆயவும் சூழலில் கூடுதல் தகுதிவாய்ந்த கணுக்களை (பொருண்மைகளை) அடையாளங்காணவும் வேண்டி பக்கத்தர வழிமுறை வரைவின் (PageRank algorithm) பயன்பாட்டின் அடிப்படையில் அமைந்த அணுகுமுறையை முன்வைக்கின்றனர். இந்நெறிமுறை ஒரு பனுவலில் உள்ள சொற்களின் எல்லாச் சாத்தியமான பொருண்மைகளையும் உருப்படுத்தம் செய்யும் மற்றும் அர்த்தமுள்ள உறவுகளுடன் பொருண்மைகளின் இணைகளை ஒன்றொடொன்று இணைக்கும் ஒரு வரைபடத்தை உருவாக்குகின்றது. உறவுகள் சொல்வலையின் உறவுகளையும் (relations from WordNet) (ஒரே உள்ளடக்குமொழியத்தைக் கொண்ட கருத்துருக்களை இணைக்கும்) சமநிலை உறவுகளையும் (coordinate relations) உட்படுத்தும். வரைபடத்திற்குப் பக்கத்தரத்தின் பயன்பாட்டிற்குப் பின்னர் சூழலில் ஒவ்வொரு சொல்லின் மிக உயர்ந்த தரம் உள்ள பொருண்மை தெரிந்தெடுக்கப்படுகின்றது.

4.5 பிற அணுகுமுறைகள்

4.5.1 சொற்பொருண்மை ஆதிக்கத்தை நிர்ணையித்தல் (Determining Word Sense Dominance)

ஒரு சொல் தரப்படுகையில் பனுவலில் அதன் பொருண்மைகளின் நிகழ்வுஎண் வினியோகம் அதிக அளவில் திரித்துக் கூறப்படுவது கவனிக்கப்பட்டுள்ளது (Kilgarriff and Rosenzweig 2000); இது சொற்பொருண்மை மயக்கநீக்கத்தின் நடத்தையைப் பாதிக்கின்றது. சொற்பொருண்மை அதிகாரத்தை நிர்ணயிக்கும் நெறிமுறை இவ்வுற்றுநோக்கு அடிப்படையில் வகை அடிப்படையிலான மயக்கநீக்கத்தை நடைமுறைபடுத்துகின்றது.

மெக்கார்த்தி (McCarthy et al 2004, 2007) அடையாளப்படுத்தப்படாத பனுவலிலிருந்து மயக்கமுள்ள சொற்களின் பொருண்மைகளைத் தானியக்கமாகத் தரப்படுத்தும் ஒரு கண்காணிக்கப்படாத நெறிமுறையை முன்மொழிந்துள்ளார். வருகைமுறை அடிப்படையில் ஒற்றுமையுள்ள அடுத்துவருபவைகள் ஒரு சொல்லின் பொருண்மையைப்பற்றி குறிப்புகளைத் தருகின்றது என்பது இவ்வணுகுமுறையின் முக்கியக்குறிப்பாகும். ஒரு இலக்குச் சொல்லுக்கு அடுத்துவருபவைகளின் ஒரு குழுமம் உள்ளது என்று எடுத்துக்கொண்டால் பொருண்மையைத் தரப்படுத்தல் அடுத்துவருபவைகளுக்கு இடையே உள்ள மற்றும் பல்பொருண்மைத்தன்மையான இலக்குச் சொல்லின் பொருண்மை விளக்கங்களுக்கு இடையே உள்ள ஒற்றுமையின் அளபைத் தகுதியாக்குவதற்குச் சமமாகும்.

4.5.2 பொருட்புல இயக்க பொருண்மை மயக்கநீக்கம் (Domain-Driven disambiguation)

பொருட்புல இயக்கப் பொருண்மை மயக்கநீக்கம் (Domain-Driven disambiguation) பொருட்புலதைப் பயன்படுத்தும் ஒரு சொற்பொருண்மை மயக்கநீக்க நெறிமுறையாகும். இலக்குச் சொல்லின் பொருண்மையானது சூழல் சொற்களின் பொருட்புலங்களுக்கும் இலக்குப் பொருண்மையின் பொருட்புலத்திற்கும் இடையேயுள்ள ஒப்பீடு அடிப்படையில் தெரிந்தெடுக்கப்படுகின்றது. இதற்குச் சொல்வலை பொருட்புல புலக்குறிப்புகள் சிறப்பாகப் பயன்படுத்தப்படுகின்றன.

முக்கியமாகப் பொருட்புலச் சொற்களின் பொருண்மை மயக்கநீக்கம் செய்யப் பொருட்புலத் தகவல்களைப் பயன்படுத்த இயலும் என்ற உண்மையின் காரணமாக இவ்வணுகுமுறை நல்ல துல்லியததைப் பெறுகின்றது. பொருட்புலத் தகவல்கள் பொருட்புல வெக்டர்களால் உருப்படுத்தம் செய்யப்படும்; அதாவது வேறுபட்ட பொருட்புலங்களிலிருந்துள்ள தகவல்களை உருப்படுத்தம் செய்யும் கூறுகளின் வெக்டார்களால் உருப்படுத்தம் செய்யப்படும்.

பொருட்புல இயக்கப் பொருண்மை மயக்கநீக்கம் மற்றும் சொற் பொருண்மை ஆதிக்க நிர்ணயம் பற்றிய ஆர்வமுள்ள விசயம் அவை மொழியியல்சார் புரிதலிலிருந்து பொருண்மை மயக்கத்தின் பொருட்புலம்சார் வகை அடிப்படையிலான மாதிரிக்கு அவற்றின் கவனக்குவிப்பை மாற்றுகிறது என்பதாகும். இவ்விலக்கு நோக்கி பொருண்மை-விழிப்புணர்வு பயன்பாடுகளைச் சாத்தியமாக்குவதன் நோக்கத்துடன் எதிர்காலத்தில் மேலும் ஆயப்படும்.

4.5.3 சொற்பொருண்மை மயக்கநீக்கத்திலிருந்து மொழி கடந்த சான்று (WSD from Cross-Lingual Evidence)

இறுதியாக மொழிபெயர்ப்புத் தகவல்களிலிருந்து கிடைக்கும் சான்றின் அடிப்படையில் பொருண்மை மயக்கநீக்கத்திற்கான ஒரு அணுகுமுறை அறிமுகப்படுத்தபடுகின்றது. இது பொருத்தமான மொழிபெயர்ப்பால் புலக்குறிப்பு செய்து இலக்குச் சொற்களின் பொருண்மை மயக்கநீக்கும் நடவடிக்கையாகும்.

ஒரு சூழலில் ஒரு சொல்லின் சாத்தியமான மொழிபெயர்ப்புகள் ஒரு துணைக் குழுமத்தின் சாத்தியமான பொருண்மைகளைக் கட்டுப்படுத்துகின்றது என்பது இவ்வணுகுமுறையின் முக்கியமான கருத்து ஆகும் (Rensnik and Yarowsky 1997, 1999). எடுத்துக்காட்டாக sentence என்ற ஆங்கிலச் சொல் சூழல் அடிப்படையில் பிரஞ்சில் peine அல்லது phrase என மொழிபெயர்க்கப்படலாம். இருப்பினும் ஒரே சொல்லின் வேறுபட்ட அர்த்தங்களுக்கு ஒரே

மொழிபெயர்ப்பு வருவது சாத்தியமானதாகும் என்பதால் இந்நெறிமுறை ஒரு முழுவதுமான பொருண்மை மயக்கநீக்கத்தை நடைமுறைப்படுத்தாது (எடுத்துக்காட்டாக *wing* என்ற சொல் ஒரு உடல் உறுப்பாகவும் ஒரு கட்டிடத்தின் உறுப்பாகவும் பொருண்மைகளை வெளிப்படுத்தும்; இவ்விரு பொருண்மைகளும் இத்தாலிய மொழியில் *ala* என்று தான் மொழிபெயர்க்கப்படுகின்றன.) ரெஸ்னிக்கும் யரோவ்ஸ்கியும் தங்களது கட்டுரையில் (Resnik and Yarowsky 1997) குறைவான எண்ணிக்கையிலான மொழிகளில் ஒரு குழுமத்தில் மொழிகளைக் கடந்து சொல்லனாக்கம் செய்யப்பட்ட பொருண்மைகள் மட்டுமே கருத்தில் கொள்ளப்படவேண்டும் என்று மொழிந்துள்ளனர். எடுத்துக்காட்டாக *table* என்பது அதன் பர்னிச்சர் பொருண்மையிலும் ஒரு மேசையின் முன்பிருக்கும் ஆட்களின் குழுவைக் குறிப்பிடும் பொருண்மையிலும் பிரஞ்சு மொழியில் *table* என்றும் இத்தாலியில் *tavola* என்றும் மொழிபெயர்க்கப்படுகின்றது. இச்சீரான பல்பொருண்மை மூன்று மொழிகளைக் கடந்து தக்கவைக்கப்படுகின்றது. இம்முன்மொழிபை நடைமுறைப்படுத்த ஐதே (Ide 2000) மொழிகளைக் கடந்து வேறாகப் பொருண்மைகளைச் சொல்லனாக்கம் செய்யும் போக்கை அடையாளம் காண ஒரு இயைபு சொல்லடைவின் (coherence index) பயன்பாட்டை குறிப்பிட்டார்.

தொடர்புள்ள படைப்புகளில் மொழிகளைக் கடந்த எடுத்துக்காட்டுகள் அடிப்படையில் பல நெறிமுறைகள் விளக்கப்பட்டுள்ளன. பிரவுண் மற்றும் பிறர் (Brown et al 1991) கண்காணிக்கப்படாத அணுகுமுறையை முன்மொழிந்தனர்; இது ஒரு இணைத் தரவுத்தொகுதியில் சொல் வரிசைப்படுத்தலை (word alignment) நடைமுறை படுத்திய பின்னர் சூழல் பண்புக்கூறுகளின் ஒரு குழுமத்திலிருந்து மிகக்கூடுதலான பண்புக்கூறுகள் அடிப்படையில் ஒரு இலக்குச் சொல்லுக்கு மிகப் பொருத்தமான மொழிபெயர்ப்பை நிர்ணயித்தது.

கேல் மற்றும் பிறர் (Gale et al 1992d) ஒரு பொருண்மை-அடையாளப்படுத்தப்பட்ட தரவுக் குழுமத்தின் தானியக்க உருவாக்கத்திற்கு இணைத் தரவுத்தொகுதியைப் பயன்படுத்தும் ஒரு நெறிமுறையை முன்மொழிந்தனர். ஒரு இலக்குச் சொல் தரப்படுகையில், மூல மொழியிலுள்ள ஒவ்வொரு வாக்கியமும் இலக்குமொழியில் அச்சொல்லின் மொழிபெயர்ப்பால் அடையாளப்படுத்தப்பட்டுள்ளது. ஒரு இயல்பான/எளிமையான பேயஸ் வகைப்படுத்தி விளையும் தரவுக் குழுமத்தால் பயிற்சி அளிக்கப்படுகின்றது மற்றும் ஒரு சொற்பொருண்மை மயக்கநீக்கச் செயலில் பயன்படுத்தப்படுகின்றது.

மிக அண்மைப்பட்ட காலத்தில் டியாப் (Diab 2003) இணைத் தரவுத்தொகுதிகளைப் பொருண்மை-அடையாளப்படுத்துவதற்குக் கண்காணிக்கப்படாத அணுகுமுறையை முன்வைத்துள்ளனர்; இது ஒரே இலக்குச் சொல்லை மொழிபெயர்க்கும் மூலச் சொற்களைக் கொத்தாக்கம் செய்கின்றது; மற்றும் ஒற்றுமையின் அளபு அடிப்படையில் பொருண்மை மயக்கநீக்கம் செய்கின்றது. இறுதியாக இந்நெறிமுறை மூலத்தரவுத்தொகுதியில் நேரும் இலக்குச் சொல்லுக்கு மிக ஒற்றுமையுள்ள அர்த்தத்தை ஒதுக்குகின்றது (மற்றும் இலக்குத் தரவுத்தொகுதியில் பொருத்தமான சொல்லுக்கு அர்த்தங்களை முன் நிறுத்துகின்றது.)

ஐதே மற்றும் பிறர் (Ide et al 2002) மற்றும் டுபிஸ் மற்றும் பிறர் (Tufis et al 2004) யூரோ சொல்வலையை (EuroWordNet) பயன்படுத்தும் ஒரு அறிவு அடிப்படையிலான அணுகுமுறையை முன்வைக்கின்றனர். இரு வரிசைப்படுத்தப்பட்ட சொற்கள் ஒரு இணைத் தரவுத்தொகுதியில் தரப்படுகையில் அவை யூரோ சொல்வலையின் இடைமொழி சொல்லடை வழியாகப் பொருத்தப்பட்ட அவ்விரு சொற்களின் ஒருபொருள்பன்மொழிகளால் பொருண்மை-அடையாளப்படுத்தப்படுகின்றது. மிக அதிக நிகழ்வெண்ணுள்ள பொருண்மை அடிப்படையானது, (most frequent sense baseline) மூல மொழியில் உள்ள சொல்லின் ஒன்றுக்கும் மேற்பட்ட பொருண்மைகள் இலக்கு மொழியில் உள்ள சொல்லின் பொருண்மைகளுடன் பொருத்தும் போது ஒரு பேக் ஆப்பாகப் (backoff) பயன்படுத்தப்படுகின்றது.

அண்மைக் கால ஆய்வுகளில் சொற்பொருண்மை மயக்கநீக்கத்திற்கு மொழிகடந்த சான்றுகளைப் பயன்படுத்தும் அணுகுமுறைள் எல்லா-சொற்கள் பொருண்மைமயக்கநீக்கத்தில் தரமான நிறைவேற்றத்தைப் பெறுகின்றது என்று கண்டுபிடிக்கப்பட்டுள்ளது (எ.கா. Ng et al. 2003), Chklovski et al. 2004; Chang and Ng 2005). இருப்பினும் இவ்வணுகுமுறைகளின் முதன்மையான சிக்கல் அறிவுப்பேறின் முட்டுக்கட்டையில் இருக்கின்றது: பல மொழிகளுக்கு இணைத் தரவுத்தொகுதிகள் குறைவாகவே உள்ளது அல்லது இல்லை எனலாம்; இதை இணைய தளங்களிலிருந்து தரவுத்தொகுதிகளைச் சேகரிப்பதால் நேர்செய்யலாம் (Resnik and Smith 2003). இச்சிக்கலைத் தீர்க்க டகான் மற்றும் இதை (Dagan and Itai 1994) என்போர் (இலக்குப் பொருண்மைகளின் குழுமமாகக் கருதப்படும்) ஒரு தொடரியல் உறவில் நேரும் பொருண்மை

மயக்கம் உள்ள எல்லாச் சாத்தியமான மொழிபெயர்ப்புகளையும் கண்டுபிடிக்க ஒரு இருமொழிய சொற்களஞ்சியத்தின்/பேரகராதியின் பயன்பாட்டை முன்மொழிகின்றனர்.

4.6 மதிப்பீட்டு நெறிமுறை (evaluation methodology)

இங்கு நாம் மதிப்பீட்டு நடவடிக்கைகள் மற்றும் பொருண்மை மயக்கநீக்க அமைப்பொழுங்குகளுக்கு மதிப்பீட்டிற்கு பயன்படுத்தப்படும் தொடக்கநிலை முறைகள் முன்வைக்கப்பட்டுள்ளன; அவைகள் தனியாகவும் பயன்பாட்டுச் சுதந்திரமாகவும் இருப்பதாகக் கருதப்படுகின்றது. இருப்பினும் சொற்பொருண்மை மயக்கநீக்கத்தின் உண்மையான நோக்கங்களில் ஒன்று தகவல் மீட்பு, இயந்திர மொழிபெயர்ப்பு போன்ற பயன்பாடுகளின் செயல்பாடுகளை மேம்படுத்துகின்றது என்பதை நிருப்பிப்பது ஆகும். பயன்பாடுகளில் உட்படுத்தப்பட்டுள்ள ஒரு தொகுதியாகச் சொற்பொருண்மை மயக்கநீக்கத்தின் மதிப்பீடு இறுதிக்கு-இறுதி மதிப்பீடு (in vivo or end-to-end evaluation) என்று குறிப்பிடப் படுகின்றது.

4.6.1 மதிப்பீட்டு நடவடிக்கைகள் (evaluation measures)

சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் மதிப்பீடு தகவல் மீட்புக் களத்திலிருந்து கடன்வாங்கப்பட்ட மதிப்பீட்டு நடவடிக்கைகளின் அடிப்படையில் பெரும்பாலும் செயல்படுகின்றது.

மேற்சொன்ன நடவடிக்கைகள் ஒரு தரப்பட்ட அர்த்ததின் விருப்பத்தேர்வுக்கு நம்பிக்கையின் ஒரு அளபு வெளியீடுக்கு ஒழுங்குமுறையின் ஆற்றலைப் பிரதிபலிக்கவில்லை என்று விவாதிக்கப்படுகின்றது.

4.6.2. தொடக்கநிலைகள் (baselines)

தொடக்க நிலைகள் வேறுபட்ட அணுகுமுறைகளின் செயன்மைகளுக்குத் தரமான நெறிமுறையாகும். இரண்டு தொடக்க நிலைகள் உள்ளன: முறையற்ற தொடக்கநிலை மற்றும் முதல் அர்த்த தொடக்கநிலை.

4.7. மதிப்பீடு: சென்ஸ்வல்/செம்வல் போட்டிகள் (senseval/semeval competitions)

வேறுபட்ட பரிசோதனைக் குழுமங்கள், பொருண்மை தெரிவடைகள் மற்றும் ஏற்றுக்கொள்ளப்பட்ட அறிவு மூலவளங்கள் காரணமாக வேறுபட்ட சொற்பொருண்மை மயக்கநீக்கத்தை ஒப்பீடு செய்வதும் மதிப்பீடு செய்வதும் மிகக்கடினமாக இருக்கின்றது. குறிப்பிட்ட மதிப்பீட்டு நடவடிக்கைகளின் நிர்வகிப்பைப் பற்றிக் கூறுவதற்கு முன் இப்பகுதியில்

இன்ஹவுஸ் (in-hourse), பெரும்பான்மையும் குறுகிய அளவு தரவுக் குழுமங்களால் மதிப்பிடப்படுகின்ற பல ஒழுங்குமுறைகள் பற்றி அறிமுகப்படுத்தப்படுகின்றன. எனவே பெரும்பாலான முன்சென்ஸ்வல் (pre-senseval) முடிவுகள் இவ்வாய்வுக்களத்தில் பின்னர் தொடர்ந்துவரும் அணுகுமுறைகளுடன் ஒப்பிடத்தக்கதல்ல.

4.7.1 சென்ஸ்வல்

(தற்போது செம்வல்/Semeval என்று மறுபெயரிடப்பட்டுள்ள) சென்ஸ்வல் என்பது 1998-இலிருந்து நடைபெறும் ஒரு அனைத்துலகச் சொற் பொருண்மை மயக்கநீக்கப் போட்டி ஆகும். வேறுபட்ட மொழிகளுக்கு எல்லா-சொற்கள் மற்றும் சொல்சார் மாதிரிக்கூறு சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகள் மற்றும் சமீபத்தில் பொருண்மையியல்சார் பங்களிப்பைப் புலக்குறிப்பு செய்தல், சொல்விளக்க சொற்பொருள் மயக்கநீக்கம், சொற்சார் இடப்பெயர்ப்பு போன்ற புதிய செயற்பாடுகளை உள்ளடக்கிய சொற்பொருண்மை மயக்கநீக்கத்தின் பல்வகையான செயல்பாடுகளுக்குச் சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் ஒப்பீட்டு மதிப்பீட்டை நடத்துவது இப்போட்டியின் நோக்கமாகும். இப்போட்டிகளின் மதிப்பீட்டிற்குச் சமர்ப்பிக்கப்படும் ஒழுங்குமுறைகள் பெரும்பாலும் வேறுபட்ட நுட்பங்களை ஒன்றிணைக்கின்றன; மற்றும் அடிக்கடி கண்காணிக்கப்பட நெறிமுறைகளையும் அறிவு அடிப்படையிலான நெறிமுறைகளையும் ஒன்றுசேர்க்கின்றன. (பயிற்சி எடுத்துக்காட்டுகளின் குறைவால் ஏற்படும் மோசமான நிறைவேற்றலைத் தவிர்க்க நெறிமுறைகள் இவ்வாறு ஒன்றிணைக்கப்படவோ ஒன்றுசேர்க்கப்படவோ செய்யப்படுகின்றன.) சென்ஸ்வல் பணிப்பட்டறைகள் சொற்பொருண்மை மயக்கநீக்கத்தின் சமீபகாலப் போக்குகளுக்கு மிக நல்ல மேற்கோள்களைக் காட்டுகின்றன. மேலும் அவை ஆய்வுச் சமூகத்திற்கு மிக மதிப்புள்ள தரவுக் குழுமங்களின் கால அடிப்படையிலான வெளியீட்டிற்கு வழிகாட்டுகின்றன.

4.7.2 சென்ஸ்வல்/செம்வல் போடிகளைப்பற்றிய கூற்றுகள்

பல காரணங்களால் மதிப்பீட்டு நடவடிக்கைகளின் ஸ்டேட்-ஆப்-ஆர்ட் ஒழுங்குமுறைகளின் நிறைவேற்றங்களை ஒப்பிடுவது மிகக் கடினமாகும். முதலாவது வேறுபட்ட அகராதிகள் பயன்படுத்தப்படுகின்றன (Senseval-1 இல் HECCTOR, Senseval-2 இல் WordNet 1.7, Senseval-3 இல் WordNet 1.7.1, Senseval-2007-இல் WordNet 2.1 மற்றும் அறைகுறை நுணுக்கமான தெரிவடைகள்). இரண்டாவது ஒழுங்குமுறைகள் பெரும்பாலும் பல

அணுகுமுறைகளை ஒருங்கிணைப்பதால் பெரும்பாலான ஒழுங்குமுறைகளில் ஒரு தனிப்பட்ட நுட்பத்தின் பங்களிப்பை மதிப்பிடுதல் கடினமாகும். மூன்றாவது கண்காணிக்கப்பட்ட ஒழுங்குமுறைகள் வேறுபட்ட தரவுத்தொகுதிகளால் பயிற்சி அளிக்கப்பட்டுள்ளன மற்றும் அறிவு அடிப்படையிலான ஒழுங்குமுறைகள் வேறுபட்ட மூலவளங்களைப் பயன்படுத்துகின்றன. இறுதியாக செம்வல்-2007 அரைகுறை நுணுக்கமான சொற்பொருண்மை மயக்கநீக்கத்திற்கு அதன் கவனத்தை மாற்றியுள்ளது.

4.8. பயன்பாடுகள்

துரதிருஷ்டவசமாக இன்றுவரைச் சொற்பொருண்மை மயக்கநீக்கம் வெளிப்படையிலான மனித மொழி தொழில்நுட்பப் பயன்பாடுகளில் உண்மையான நன்மைகளை இதுவரை நிரூபிக்கவில்லை என்று கூறலாம். இருப்பினும் இறுதியிலிருந்து-இறுதி பயன்பாடுகளின் குறைவு சொற்பொருண்மை மயக்கநீக்கத்தின் தற்போதைய நிறைவேற்றத்தின் விளைவாகும்; மற்றும் அது கூடுதல் துல்லியமான பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகளையும் எதிர்காலத்தில் பொருண்மையியல் அடிப்படையிலான இயற்கை மொழி ஆய்வுப் பயன்பாடுகளின் சாத்தியத்தையும் தடைசெய்யாது. ஒரு மேம்பட்ட துல்லியம் புதுமையான நெறிமுறைகளிலிருந்து கிடைப்பதுடன் பொருண்மை மயக்கநீக்கச் செயல்பாடுகளின் வேறுபட்ட பின்னணி அமைப்புகளிலிருந்தும் கிடைக்கின்றது.

இங்கு நாம் சொற்பொருண்மை மயக்கநீக்கத்திலிருந்து கிடைக்கப்பெறும் மற்றும் பரிசோதனைக்கு உட்படுத்தபட்ட பல உண்மை உலகப் பயன்பாடுகளைப் பற்றிய சுருக்கம் இங்கு கூறப்படும்.

4.8.1 தகவல் மீட்பு (Information Retrieval (IR)

ஒரு பயன்பாட்டாளரின் கேள்விக்குப் பொருத்தமற்ற ஆவணங்களை நீக்குவதற்கு ஸ்டேட்-ஆஃப்-ஆர்ட் (State-of-art) தேடல் இயந்திரங்கள் வெளிப்படையிலான பொருண்மையியலைப் பயன்படுத்தவில்லை. கேள்விச் சொற்களின் ஒரு சாத்தியமான சொற்பொருண்மை மயக்கநீக்கத்துடன் கூடிய ஆவண அடிப்படையின் ஒரு துல்லியமான பொருண்மை மயக்கநீக்கம் வேறுபட்ட பொருண்மைகளைப் பயன்படுத்தும் ஒரே சொற்களைக் கொண்டிருக்கும் ஆவணங்களை நீக்கவும் ஒரே சொற்களை வெளிப்படுத்தும் ஆவணங்களை மீட்கவும் அனுமதிக்கும். (இது மீள் அழைப்பை/recall அதிகரிக்கும்.)

தகவல் மீட்புகளுக்குச் சொற் பொருண்மை மயக்க நிக்கத்தின் பங்களிப்பின் பெரும்பாலான தொடக்ககாலச் செயல்பாடுகள், நிறைவேற்ற முன்னேற்றமில்லாமையில் விளைந்தது (எ.கா. Salton 1968; Salton and McGill 1983; Krovetz and Croft 1992; Voorhees 1993; Sanderson 2000). க்ரொவெட்ஸ் மற்றும் க்ரொஃப்ட் (Krovetz and Croft 1992) மற்றும் சாண்டர்சன் (Sanderson 2000) என்போர் கேள்விச் சொற்களின் ஒரு குறைந்த சதவிகிதம் தான் அவற்றின் கூடுதல் நிகழ்வெண் (முக்கியமான) பொருண்மையில் பயன்படுத்தப்படவில்லை என்று காட்டினர்; இது சொற்பொருண்மை மயக்கநீக்கம் கூடுதல் நிகழ்வெண்ணுள்ள சொற்களைக் காட்டிலும் அரிய சொற்களில் தான் மிகத் துல்லியமானது என்று காட்டுகின்றது. சாண்டர்சன் (Sanderson 1994) மிகுந்த எண்ணிக்கையிலான சொற்களுடன் கேள்விகள் இருக்கும்போது சொற்பொருண்மை மயக்கநீக்கம் தகவல் மீட்புக்கு நன்மை தராது என்ற முடிவுக்கு வருகின்றார். அவர் மிகச் சிறிய கேள்விகள் மிகுந்த பொருண்மை மயக்கத் திறன் உள்ளதாக இருக்கவியலும் எனக் குறிப்பிடுகின்றார்.

4.8.2 தகவல் பிரித்தெடுப்பு (information extraction)

குறிப்பிட்ட பொருட்புலங்களில் கருத்துருக்களின் குறிப்பிட்ட நேர்வுகளுக்கு இடையே வேறுபாடுகளைக் காண்பது ஆர்வமுள்ளதாகும்: எடுத்துக்காட்டாக மருத்துவப் பொருட்புலத்தில் நாம் ஒரு பனுவலைக் கடந்து எல்லா வகையில்படும் மருந்துகளை அடையாளம் காண்பதில் விருப்பம் உள்ளவர்களாய் இருப்போம்; ஆனால் உயிரியல்தகவலியலில் (bioinformatics) நாம் உயிரணுக்கள் மற்றும் புரதச்சத்துக்கள் இவற்றைப் பெயரிடுவதில் உள்ள பொருண்மை மயக்கநீக்கத்தைத் தீர்க்க விழைவோம். பெயரிடப்பட்ட-இருப்புப்பொருள்ளை அறிதல் (Named-entity recongnition (NER)), தலைப்பனெழுது விரிவு (acronym expansion) (எ.கா. MP = Member of Parliament or Military Police) என்பனவற்றைச் சொற்பொருண்மை மயக்கநீக்கச் சிக்கல்களாகக் கருதவியலும்; இருப்பினும் இது ஒத்தறி அடிப்படையில் ஒரு புதிய களமாகும் (Dill et al 2003).

ஜாக்மின் மற்றும் பிறர் (Jacquemin et al 2002) சொல், பொருட்புலம் மற்றும் தொடரியல் மற்றும் பொருண்மையியல் நிலையில் பொருண்மைமயக்கநீக்க விதிகளின் பயன்பாட்டைக் கொண்டிருக்கும் அகராதி அடிப்படையிலான நெறிமுறையை முன்வைக்கின்றனர். மாலின் மற்றும் பிறர் (Malin et al 2005) பெயரிடப்பட்ட இருப்புப்பொருள்களின் பொருண்மை மயக்கநீக்கத்தைத் தீர்க்க ஒழுங்கற்ற இயக்கங்கள் (random walks) அடிப்படையிலான ஒரு

தொடர்புப் பகுப்பாய்வு நெறிமுறையின் (link analysis method) பயன்பாட்டை முன்மொழிகின்றார். ஹாசன் மற்றும் பிறர் (Hassan et al 2006) நேர்வுகளின் ஒரு குழுமத்தின் மீது அவற்றின்/தமது தாக்கம் அடிப்படையிலான இருப்புப்பொருள் பிரிப்பு ஒழுங்குமுறைகளை (entity extraction patterns) அளக்கப் பகுதி கண்காணிக்கப்பட்ட நடையில் ஒரு இணைப்புப் பகுப்பாய்வு வழிமுறைவரைவைப் (link analysis algorithim) பயன்படுத்தினார். இறுதியாக சியரமிதா மற்றும் ஆல்டன் (Ciaramita and Altun 2006) சொல் வலை ஒருபொருள்பன்மொழியக் குழுமங்களின் ஒரு கட்டுப்படுத்தப்பட்ட குழுமத்திலிருந்து தெரிந்தெடுக்கப்பட்ட ஒரு வகுப்பை ஒதுக்கும் ஒரு உயர் பொருண்மை அடையாளப்படுத்தியின் பயன்பாட்டை முன்மொழிகின்றார். ஹிட்டன் மார்கோவ் மாதிரிகளால் கோர்வை புலக்குறிப்பு செய்வது (sequence labeling) அடிபடையிலான இவ்வணுகுமுறை ஒரு கற்றல் நடவடிக்கையை வேண்டுகின்றது.

4.8.3 இயந்திர மொழிபெயர்ப்பு (Machine Translation)

ஒரு சூழலில் ஒரு சொல்லின் சரியான மொழிபெயர்ப்பின் தானியங்கு அடையாளம் காணல் அதாவது இயந்திர மொழிபெயர்ப்பு ஒரு மிக்க கடிமான வேலையாகும். பனுவல்களின் பொருண்மை மயக்கநீக்கம் மொழிபெயர்ப்பு ஒழுங்குமுறைகளைச் சரியான சொற்களைத் தெரிந்தெடுக்க உதவவேண்டும் என்ற உள்ளுணர்வுக் கருத்து அடிப்படையில் இயந்திர மொழிபெயர்ப்பைச் சாத்தியமாக்க வேண்டி தீர்க்கப்படவேண்டிய முக்கியமான செயலாகச் சொற்பொருண்மை மயக்கநீக்கம் வரலாற்று அடிப்படையில் ஏற்றுக்கொள்ளபட்டுள்ளது. சூழலைப் பொறுத்துச் சொற்கள் முற்றிலும் வேறுபட்ட மொழிபெயர்ப்புகளைப் பெறும். எடுத்துக்காட்டாக, line என்ற ஆங்கிலச் சொல் இத்தாலிய மொழியில் linea, riga, verso, filo, corda என மொழிபெயர்ப்பு செய்யப்படலாம். துரதிருஷ்டவசமாகச் சொற்பொருண்மை மயக்க நீக்கம் எதிர்பார்த்ததைவிட மிகக் கடினமானது என நாம் பல ஆண்டு ஒப்பீட்டு மதிப்பீடுகளின் பின்னர் அறிந்துகொண்டோம். முன்னர் கூறியபடி சொற்பொருண்மை மயக்கநீக்கத்தில் 1960களில் ஏற்பட்ட ஆரம்பகாலத் தோல்வி இயந்திர மொழிபெயர்ப்புக் களத்தை மிக மோசமான நிலைக்குக் கொண்டு சென்றது. தற்போது இதற்கு முரணாகச் சொற்பொருண்மை மயக்க நீக்கம் இயந்திர மொழிபெயர்ப்புக்கும் உதவும் என்ற அறிகுறி கிடைத்துள்ளது. கார்பாட் மற்றும் வூ (Carpuat and Wu 2005) என்பவர்கள் சொற்பொருண்மை மயக்கநீக்கத்தைத் தற்கால இயந்திர மொழிபெயர்ப்புப் பயன்படுகளில் இணைக்க இயலும் என்று குறிப்பிடுகின்றனர்; டாகன் மற்றும்

இதை (Dargan and Itai 1994), விக்ரே மற்றும் பிறர் (Vickrey et al 2005) சொற்பொருண்மை மயக்க நீக்கத்தின் சரியான பயன்பாடு மொழிபெயர்ப்பு நிறைவேற்றத்தின் அதிகரிப்புக்குக் கொண்டு செல்லும் என்று காட்டுகின்றனர்.

மிகச் சமீப காலத்தில் கார்பாட் மற்றும் வூ (Carpuat and Wu 2007) மற்றும் சான் மற்றும் பிறர் (Chan et al 2007a) சொற்பொருண்மை மயக்கநீக்கம் இயந்திர மொழிபெயர்ப்பை மேம்படுத்த உதவ இயலும் என்று காட்டியுள்ளனர். இவ்வாய்வுகளில் மிகப் பெருத்தமான மொழிபெயர்ப்புத் தொடரைத் தெரிந்தெடுக்க அனுமதிக்கும் சொற்பொருண்மை மயக்க நீக்க மாதிரிகள் ஏற்றுக்கொள்ளப்பட்டு முன்-வரையறை விளக்கத் தெரிவடைகள் (predefined sense inventories) கைவிடப்பட்டுள்ளன. இருப்பினும் இம்முடிவுகள் ஆய்வுக்களத்தை இயந்திர மொழிபெயர்ப்பின் வெற்றிக்கு மரபு சொற்பொருண்மை மயக்கநீக்கத்தின் பங்களிப்புமீதான கருதுகோள்களுக்குத் திறந்து விட்டுள்ளது.

4.8.4 பொருளடக்க ஆய்வு (content analysis)

கருத்து, மையக்கருத்து போன்றவற்றின் அடிப்படையில் ஒரு பனுவலின் பொதுவான பொருளடக்கத்தின் பகுப்பாய்வு சொற்பொருண்மை மயக்க நீக்கத்தின் பயன்பாட்டால் நிச்சயமாக நன்மை பெறலாம். எடுத்துக்காட்டாக, பிளாக்குகளை (blogs) வகைப்படுத்துவது இணையதள சமூகத்திற்கு இடையே கூடுதல் ஆர்வத்தைப் பெற்றுள்ளது: பிளாக்குகள் மிக விரைவாக வளர்வதால், அவற்றை வகைப்படுத்தவும், அவற்றின் முக்கியமான தலைப்புகளை நிர்ணயிக்கவும் மற்றும் பிளாக்குகளுக்கு இடையிலும் தனி பிளாக் போஸ்டுகளுக்கு (single blog posts) இடையிலும் உள்ள பொருத்தமான தொடர்புகளை அடையாளம் காணவும் எளிதான திறமையான வழி வேண்டும். ஆய்வின் இரண்டாவது தொடர்புள்ள களம் சமுதாய வலைப்பின்னல் பகுப்பாய்வு (social network analysis) ஆகும்; இது இணைய வலையின் சமீபகால மதிப்பீடுகளால் கூடுதல் செயலுக்கம் உடையதாய் மாறுகின்றது.

4.8.5 சொல் பகுப்பாய்வு (word processing)

சொல் பகுப்பாய்வு இயற்கை மொழி ஆய்வின் பொருத்தமான பயன்பாடாகும்; இதன் முக்கியத்துவம் நீண்ட காலமாகவே அறியப்பட்டுள்ளது (Churuch and Rau 1995). சொற்பொருண்மை மயக்கநீக்கம் ஒரு சொல்லின் எழுத்துக்கூட்டலைச் சரிசெய்தல், வேற்றுமை மாற்றம் செய்தல், பொருத்தமான ஒலிக்குறியீடுகளைச் செருகுதல் போன்றவற்றிற்கு உதவும்.

4.8.6 அகராதியியல் (lexicography)

சொற்பொருண்மை மயக்கநீக்கமும் அகராதியியலும் ஒன்றிலிருந்து ஒன்று நன்மை பெறும்: சொற்பொருண்மை மயக்கநீக்கம் அனுபவவாதப் பொருண்மைக் குழுமங்களையும் புதிய அல்லது ஏற்கனவே இருக்கிற அர்த்தங்களின் சூழலின் புள்ளியியல் அடிபடையில் சிறப்பான அடையாளங் காட்டிகளையும் தந்துதவ இயலும். மேலும் சொற்பொருண்மை மயக்கநீக்கம் இயந்திரம் படிக்கவியலும் அகராதிகளிலிருந்து சொற்பொருண்மை வலையமைப்புகளை உருவாக்க உதவ இயலும் (Richardson et al 1998). இதற்குப் பதிலாக ஒரு அகராதி இயலார் சொற்பொருண்மை மயக்கநீக்கத்திற்கு உதவும் பொருண்மைத் தெரிவடைகளையும் பொருண்மை-அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியையும் தருகின்றனர்.

4.8.7 பொருண்மை வலை (semantic web)

இறுதியாக, பொருண்மை வலை காட்சி (semantic web vision) (Berners-Lee et al. 2001) முன்னர் கூறிய பயன்படுகளிலிருந்து நன்மை பெற இயலும்; ஏனென்றால் அது (வலை/Web)ஆவணங்களின் பொருண்மையியலைக் கையாளுவதற்காக மரபுரிமையாகப் பெருட்புலம் அடிப்படையிலான மற்றும் எல்லையற்ற பொருண்மை மயக்க நீக்கத்தை வேண்டுகின்றது மற்றும் ஒழுங்குமுறைகளுக்கும் மூலப்பொருண்மையியல் ஆய்வுகளுக்கும் (ontology) பயன்பாட்டாளர்களுக்கும் இடையில் ஊட்டாடத்தைச் சாத்தியமாக்குகின்றது. மூலப்பொருண்மையியல் ஆய்வு கற்றல், பொருட்புல வகைப்பாட்டியலை உருவாக்குதல் (Navigali et al 2003; Navigli and Velardi 2004; Cimianao 3006), சொற்பொருண்மை மயக்கநீக்கம் மற்றும் மிகப் பெரிய அளவிலான பொருண்மை வலைப்பின்னல்களை மேம்படுத்துதல் (Navigli and Velardi 2005; Pennacchiotti and Pantel 2006; Snow et al 2006) போன்ற இணையவலை சார்ந்த ஆய்வுக் களங்களில் பயன்படுத்தப்படுகின்றது.

4.9 திறந்த சிக்கல்களும் எதிர்காலத் திசைகளும்

இப்பகுதியில் முக்கியமான திறந்த சிக்கல்களும் சொற்பொருண்மை மயக்கநீக்கக் களத்தில் எதிர்காலச் திசைகளும் சுருக்கமாகத் தரப்படும்.

4.9.1 சொற் பொருண்மையின் உருப்படுத்தம்

சொற்பொருண்மை மயக்க நீக்கத்தின் அடிப்படையான சிக்கல் எவ்வாறு சொற்பொருண்மையை உருப்படுத்தம் செய்வது என்பதன் விருப்பத் தேர்வாகும். ஒரு நோக்கில்

சொற்பொருண்மை மயக்கநீக்க ஒழுங்கு முறைகளின் புறவயமான மதிப்பீடுக்கு அர்த்தங்களைப் பட்டியலிடுகின்ற சொற்களஞ்சியங்கள்/பேரகராதிகள் மிகவும் திறமையாகச் செயல்படக்கூடிய அணுகுமுறையாகத் தோன்றுகின்றது. மாறாகக் கண்காணிக்கப்படாத வழிமுறைவரைவு மிக எளிதாக மதிப்பிடக்கூடியதாய் இருக்கின்றது, அதாவது இறுதியிலிருந்து-இறுதிப் பயன்பாடுகளுக்குத் திறமை வாய்ந்ததாய் இருக்கின்றது. இதன் அடிப்படையில் Semval-2007-இல் நிகழ்த்தப்பெற்ற மொழிகடந்த தகவல் மீட்பிலும் சொல் இடப்பெயர்ச்சியிலும் சொற்பொருண்மை மயக்கநீக்கம் போன்ற செயல்பாடுகளின் மதிப்பீடு வேறுபடுத்தப்பட்ட பொருண்மைத் தெரிவடைகளின் உண்மையான தேவையை வெளிப்படுத்துகின்றது.

பொருண்மையைப் பட்டியலிடும் அணுகுமுறையின் பரந்த பயன்பாடின் விளைவாகப் பொருண்மைகளை எவ்வாறு பிரிப்பது என்ற சிக்கல் உடனடியாக எழுகின்றது (Ide and Véronis 1998). பல ஆய்வாளர்கள் (e.g. Wilks and Slator 1989, Fellaum et al. 2001, Palmer et al. 2004), Ide and Wilks 2006) பெரும்பாலான அகராதிகளில் பொருண்மைப் பிரிவுகள் (sense divisions) பெரும்பாலான இயற்கை மொழி ஆய்வுப் பயன்பாடுகளின் தேவைகளுக்கு அதிகமாகக் கூடுதல் நுண்மையாக்கம் செய்யப்பட்டதாகும். இது முக்கியமாக இயற்கை மொழி ஆய்வுச் சமூகத்தால் பயன்படுத்தப்படும் சொல்வலைகளுக்கும் பொருந்தும்.

போதுமான அளவு நுண்மையை நிலைநிறுத்துவதன் நோக்கங்களில் ஒன்று தற்கால முன்னேற்றம் தழுவிய நிலையில் உள்ள மிக நுண்ணுக்கமான பொருண்மை மயக்கநீக்க ஒழுங்குமுறைகள் ~70% துல்லிய எல்லையை மிஞ்சுவதாகும். இது இன்றைய அளவிலும் ஒரு திறந்த சிக்கலாக இருந்தாலும் குறிப்பிட்ட பயன்பாடுகளின் தேவைக்குத் தக்கவாறு நுண்மையின் வேறுபட்ட நிலைகளைக் கண்டுபிடிக்கும் முயற்சிகள் பல மேற்கொள்ளப்பட்டன மற்றும் மேற்கொள்ளப்பட்டு வருகின்றன. இவற்றில் சில பொருண்மை கொத்தாக்கம் (sense clustering) மற்றும் சொற் பொருண்மை தூண்டல் (word sense induction) என்பனவாகும். ஒரே சொல்லின் பொருண்மைகளுக்கு இடையில் உள்ள உறவுகளின் வலுவைத் தரப்படுத்த இயலும் வழிமுறை வரவுகளின் ஒரு ஆர்வமுள்ள பண்புக்கூறு என்னவென்றால் பொருண்மைத் தெரிவடையின் நுண்மையை (granaluratity of the sense inventory) கையில் உள்ள குறிப்பிட்ட பயன்பாட்டிற்காக மாற்றவியலும்.

Semeval-2007-இல் நடத்தப்பட்ட கரடுமுரடாக நுண்மையாக்கம் செய்யப்பட்ட சொல்சார் மாதிரி (coarse-grained lexical sample) மற்றும் எல்லா-சொற்கள் சொற்பொருண்மை மயக்கநீக்கம் (all-words WSD) என்பனவற்றின் செயல்பாடுகளும் நுண்மைச் சிக்கலைக் கையாளுவதைக் குறிக்கோளாகக் கொண்டதாய் அமைந்தது. இவ்விரு செயல்பாடுகளும் முறையே ஹொவல் மற்றும் பிறர் (Hovy et al 2006) மற்றும் நவிக்லி (Navigli 2006c) என்பவர்களின் ஆய்வுகளின் அடிப்படையில் அமைந்தது. ஆன்டோ நொட்ஸ் திட்டத்தின் (OntoNotes project) சூழலில் ஹொவி மற்றும் பிறர் ஒமேகா ஆன்டாலஜிக்காக (Omega Ontology) (Philpot et al. 2005) கரடுமுரடான பொருண்மைகளை உருவாக்கினர்; சொல்வலையின் பொருண்மை தெரிவடையில் தொடங்கி சொற்பொருண்மை அடையாளப்படுத்தும் செயல்பாட்டில் அடையாளப்படுத்துபவர்க்கு இடையில் 90% உடன்பாடு கிடைக்கும் வரை பொருண்மைகளைப் மீண்டும் மீண்டும் பிரித்து இது செய்யப்பட்டது. இதற்கு முரணாக நவிக்லி (Navigli 2006c) ஆக்ஸ்போர்டு ஆங்கில அகராதியில் உள்ள (Oxfore dictionary of English) பொருண்மைப் பதிவுகளுக்குத் (sense entries) தானியக்கப் பொருத்தம் வழி சொல்வலை பொருண்மைக் கொத்துகளை (WordNet sense clusters) உருவாக்கினார்.

நுண்மைச் சிக்கலின் தெளிவான நிலை ஐதே மற்றும் வில்க்ஸ் (Ide and Wilks 2006) என்பவர்களால் எடுத்துக்கொள்ளப்பட்டது; அவர்கள் பயன்பாடுகளுக்குத் தேவைப்படும் பொருண்மை வேறுபாடுகளின் நிலை பல்பொருளொருமொழிகளுக்கு/ஒப்புருமொழிகளுக்கு (homonyms) ஓரளவுக்குப் பொருந்தும் என்று குறிப்பிடுகின்றனர்; மனிதரால் உண்மையில் பல்பொருளொருமொழிளாகக்/ஒப்புருமொழிகளாகக் கருத்தப்படும் சொற்பிறப்பியல் அடிப்படையில் உறவுள்ள பொருண்மைகள் இதற்கு விதிவிலக்காகும்.

4.9.2 அறிவுப் பேறு நெருக்கடி

வெளிப்படையாக எல்லாச் சொற்பொருண்மை மயக்கநீக்க நெறிமுறைகளும் தரவுத் தொகுதிகள் அல்லது அகராதிகள் என்ற அறிவு மூலத்தைச் சார்ந்திருக்கின்றன. எனவே அறிவுப் பேறு நெருக்கடி (knowledge acquisition bottleneck) (என்று அழைக்கப்படுவது) சொற்பொருண்மை மயக்கநீக்கத்தின் முக்கியமான விவாத விசங்களில் ஒன்றாகும். இச்சிக்கலை எளிதாக்கும் பல நுட்பங்கள் முன்னர் விளக்கப்பட்டது: ஈடேற்றம்/bootstrapping மற்றும் ஊக்கமான கற்றல் (active learning), பயிற்சித் தரவுத்தொகுதியின் தானியக்கப் பேறு (automatic

acquisition of training data), மொழிகடந்த தகவலின் பயன்பாடு (the use of cross lingual information) போன்றன. இங்கு நாம் அறிவுப்பேறு நெருக்கடியைத் தளர்த்துவதை நோக்கமாகக்கொண்ட நடைமுறைகளை விளக்கிவாதிடுவோம்: அறிவு மூலவளங்களின் தானியக்கச் செறிவூட்டல் (automatic enrichment of knowledge resources), குறிப்பாக இயந்திரம் படிக்கவியலும் அகராதிகள் (machine readable dictionaries) மற்றும் கணினிசார் சொற்களஞ்சியங்கள்/பேரகராதிகள் (computational lexicons).

அறிவுச்செறிவூட்டல் (knowledge enrichment) (Amsler [1980] and Litkowski [1978]) என்பவர்களின் ஆய்வுக்களுக்குக் கொண்டுசெல்லும் வரையறை விளக்கங்களிலிருந்து தகவலைப் பிரித்தெடுக்கும் உள்ளுணர்வு அணுகுமுறை (எ.கா. (e.g., Chodorow et al. [1985]; Rigau et al.[1998]). அணுகுமுறை மூன்று நடைமுறைகள் அடிப்படையிலானது: பொதுவினங்களைப் (genus) பெறுவதற்கான வரையறை விளக்கப் பகுத்துகுறித்தல் (அதாவது உள்ளடக்குமொழி) கருத்துரு; பொதுவினப் பொருண்மைமயக்கநீக்கம் (genus disambiguation); வகைப்பாட்டியல் உருவாக்கம் (taxonomy construction).

சொல்வலை போன்று கிடைக்கின்ற மூலவளங்களைப் புதிய பொருண்மை உறவுகளில் சேர்ந்துவருகைகள் மற்றும் உறவு முத்தொகுதிகள் (relation triples) பொருண்மை மயக்கநீக்கம் செய்யப்பட வேண்டும் (உறவுகளின் மூலப்பொருண்மையாக்கம்/ *ontologization* of relations) எ.கா. (*car*$_n$, *driver*$_n$) என்பதை (*car*1$_n$, *driver*1$_n$) ஆகமாறுதல். உறவு முத்தொகுதிகளின் பொருண்மை மயக்கநீக்கத்திற்குக் கண்காணிக்கப்பட்ட இயந்திரம் கற்றல் அணுகுமுறைகளும் பயன்படுத்தப்பட்டன [Girju et al. 2003].

இறுதியாக நாம் அறிவுப்பேறு நெருக்கடியை நீக்க இரண்டு மனிதச்செயல் முயற்சிகளைக் கூறலாம். முதல் முயற்சி Open MindWord Expert [Chklovski and Mihalcea 2002] என்று அழைக்கப்படும் அறிவுப் பேறுக்கான ஒன்றிணைந்த தளம்; இதில் இணையதளத்தில் மனித ஆர்வலர்கள் சூழலில் சொற்களுக்கு அர்த்தத்தை அடையாளப்படுத்த கேட்டுக்கொள்ளப்படுவார்கள். இவ்வணுகுமுறை இணையதள அடையாளப்படுத்துபவர்களுக்கு இடையில் உள்ள உடன்பாட்டைச் சார்ந்திருக்கின்றது. ஒரு இலக்குச் சொல் நேர்வுக்குக் கூடுதல் சாத்தியமான அர்த்த ஒதுக்கீடை உறுதிசெய்ய ஒரு பரந்த உடன்பாடு பயன்படுத்தப்படுகின்றது. இரண்டாவது முயற்சி தற்போது பிரின்ஸ்டனில் நடக்கும் சொல்வலைபிளஸ் திட்டத்தில்

(WordNetPlus project [Boyd-Graber et al. 2006]) .பொருண்மை அடையாளப்படுத்தப்பட்ட அர்த்த விளக்கங்கள் மற்றும் கருத்துரு அழைப்பு (concept evocation) (எ.கா. *egg-bacon*, *yell-voice*, etc.) அடிப்படையில் பொருண்மை உறவுகளின் பகுதி தானியக்கச் சேர்க்கை இவற்றால் சொல்வலையின் செறிவூட்டைக் கருத்தில்கொள்ளும்.

அண்மைக்காலத்திய மிகப் பெரிய அளவிலான அர்த்தப் பேறு மற்றும் செறிவூட்டல் பரந்த செயலெல்லையையும் துல்லியமான சொற்பொருண்மை மயக்கநீக்கத்தையும் சாத்தியமாக்கும்.

4.9.3 பொருட்புலம் சார் சொற்பொருண்மை மயக்கநீக்கம்

பயன்பாடுகளில் சொற்பொருண்மை மயக்கநீக்கத்தின் வெற்றிகரமான உபயோகம் இவ்வாய்வுக்களத்தின் முக்கியமான நோக்கங்களில் ஒன்றாகும். பயன்பாடுகள் யாவும் பெரும்பாலும் விருப்பமுள்ள ஒரு குறிப்பிட்ட பொருட்புலத்தின் மீது கவனக்குவிப்பு செய்தன. இருப்பினும் பொருட்புலம் சார் பொருண்மை மயக்கநீக்கத்திற்குச் சிறிதளவு கவனம் தான் தரப்பட்டது. முக்கியமான கருதுகோள் விருப்பமுள்ள பொருட்புலத்தின் அறிவு குறிப்பிட்ட பொருட்புலச் சூழலில் சொற்களின் பொருண்மை மயக்கநீக்கம் செய்ய உதவ இயலும் என்பதாகும். முக்கியமான அர்த்தத்தின் கண்டுபிடிப்பு மீதான ஆய்வுப்பணிகள் மற்றும் பொருபுல இயக்க பொருண்மை மயக்கநீக்கம் (domain-driven disambiguation) மற்றும் பொருட்புல இயைவிப்பு (domain tuning), அதாவது இலக்குப் பொருட்புலத்திற்கு அதிகப் பொருத்தமான அர்த்தங்களின் தானியியக்கத் தெரிவு [Basili et al. 1997; Cucchiarelli and Velardi 1998; Buitelaar and Sacaleanu 2001] என்பன இந்த திசையில் செல்லும்.

பொருட்புலம் அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் பொருட்புலம் சார் பயன்பாடுகளின் அதிகரிக்கும் தேவையால் தீர்மானிக்கப்படுகின்றது; எடுத்துக்காட்டாக உயிரியல் மருத்துவம், கணினி அறிவியல், சுற்றுலா மற்றும் பிற. மேலும் பொருண்மை வலை காட்சி (semantic Web vision) குறிப்பிட்ட பொருட்புலம் சார்ந்த மூலப்பொருண்மையியல் ஆய்வுகளைக் (domain-specific ontologies) கையாளுவதற்குத் திறமையை வேண்டுகின்றது. எனவே குறிப்பிட்ட அறிவுக் களங்களில் ஆய்வுப்பணிபுரியும் திறமை பொருட்புலம் சார் பயன்பாடுகளின் வெற்றிக்கு அதிக அளவில் முக்கியமாகும்.

4.10 முடிவுரை

இவ்வியலில் பொருண்மைமயக்கநீக்கம் பற்றிய பரந்துபட்ட ஆய்வு விளக்கங்களும் அதன் செயல்பாடுகளும் அதற்கான முயற்சிகளும் அதில் ஏற்பட்டுள்ள வெற்றி தோல்விகளும் தடைகளும் அதை நேரிடும் நடைமுறைத் திட்டங்களும் அதன் முன்னேற்றப் பாதைகளும் அண்மைக்காலத்தில் அதன் நிலைமைகளும் தரப்பட்டுள்ளன. சொற்பொருண்மை மயக்கநீக்கத்தின் ஆய்வு 1950களிலிருந்து நடந்துகொண்டிருக்கின்றது. சொற்பொண்மை மயக்கநீக்கம் கடினமான ஆய்வுப்பணியாகும். இது மொழியின் முழுமையான கலவைத் தன்மைகளைக் கையாளுகின்றது. அமைப்பாக்கம் செய்யப்படாத மூலப் பனுவல்களிலிருந்து பொருண்மை அமைப்பை அடையாளம் காண்பதைக் குறிக்கோளாகக் கொண்டுள்ளது.

சொற்பொருண்மை மயக்கநீக்கத்தின் கடினம் எடுத்துக்கொள்ளப்பட்ட அர்த்தங்களின் நுணுக்கத்தைச் சார்ந்திருக்கின்றது. யரோவ்ஸ்கி மற்றும் பிறர் (Yarowsky [1995] and Stevenson and Wilks [2001]) 95%க்கு அண்மையான அல்லது அதற்கு அதிகமான துல்லியம் ஒருசொல்போலிகளின் பொருண்மை மயக்கநீக்கத்தில் பெற இயலும் என்று காட்டியுள்ளனர். ஆனால் பல்பொருண்மையின் பொதுவான கருத்துச்சாயலுக்கு வருகையில் சிக்கல் கடினமாகின்றது; இங்கு நுண்மையானது, பொண்மைமயக்கநீக்க ஒழுங்குமுறைகளின் செயல்பாட்டிலும் மனித அடையாளப்படுத்துபவர்களுக்கு இடையில் உடன்பாட்டிலும் வேறுபாட்டை உருவாக்குகின்றது.

கண்காணிக்கப்பட்ட நெறிமுறைகள் பிற நெறிமுறைகளைக் காட்டிலும் எந்தவித ஐயமுமின்றி நன்றாகச் செயலாற்றுகின்றது. இருப்பினும் வேறுபட்ட பொருட்புலங்களுக்கு, மொழிகளுக்கு, பணிகளுக்கு அதிக அளவிலான பயிற்சித் தரவுத்தொகுதிகள் கிடைப்பதைச் சார்ந்திருப்பது என்பது சாத்தியமான அனுமானம் அல்ல. Ng [1997] மிக உயர்ந்த துல்லியமான பரந்த செயலெலையைக் கொண்ட பொருண்மைமயக்கநீக்க ஒழுங்குமுறையைப் பெற/உருவாக்க 3.2 மில்லியன் அர்த்தம் அடையாளப்படுத்தப்பட்ட சொற்கள் தேவை என்று நிர்ணையித்துள்ளார்.

மாறாக அறிவு அடிப்படையிலான அணுகுமுறைகள் பலகாரணங்களால் சிறிய-மத்திய செயல்பாடுகளுக்கு மிகவும் நம்பகமானவை: முதலாவது கூடுதல் அறிவு இருந்தால்/கிடைத்தால் கூடுதல் செயல்பாடு கிடைக்கும். [Cuadros and Rigau 2006; Navigli and Lapata 2007]; இரண்டாவது அவர்கள் சார்ந்திருக்கும் மூலவளங்கள் ஏறுமுகமாகச் செறிவூட்டலைப் பெறுகின்றன (எடுத்துக்காட்டாக சொல்வலைகள் சொல்வலைபிளஸ் என்பன); மூன்றாவது

பொருண்மை வலையின் பயன்பாடுகள் ஆற்றல் மிக்கப் பொருட்புல மூலப்பொருண்மையியல் ஆய்வுகளின் ஆற்றலைப் பயன்படுத்தும் மற்றும் பயன்படுத்துபவர்கள், தொழில்கள், மற்றும் ஒழுங்குமுறைகள் இவற்றிற்கிடையில் பொருண்மை ஊட்டாட்த் திறனை அறிவுச் செழுமிய நெறிமுறைகள் வேண்டுகின்றன.

இயல் 5

தமிழில் சொற்பொருள் மயக்கநீக்கத்திற்கான நெறிமுறைகள்

5.0 முன்னுரை

ஒரு வாக்கியம் அதன் அமைப்பு அடிப்படையிலும் அதில் வரும் சொற்கள் அடிப்படையிலும் பொருள்கோள் செய்யப்படுகின்றது. ஒரு வாக்கியம் பல வகைப் பகுத்துக் குறிப்புகளைப் (different types of parsing) பெறுகையில் அவ்வாக்கியம் பல பொருள்கோள்களைப் பெறும். மட்டுமன்றி வாக்கியத்தில் வரும் சொற்கள் பல்பொருண்மைத் தன்மை/பல்பொருளொருமொழியத் தன்மை உடையதாய் இருந்தாலும் சொற்போலியாக இருந்தாலும் வாக்கியம் வேறுபட்ட பொருள்கோள்களைப் பெறும். தமிழ் வாக்கியங்கள் எவ்வாறு அமைப்புப் பொருண்மை அடிப்படையிலும் சொற்பொருண்மை அடிப்படையிலும் பொருண்மை மயக்கம் அடைகின்றன என்பதை இரண்டாம் இயலில் பார்த்தோம். அதை இயந்திர மொழிபெயர்ப்பு போன்ற செயல்பாடுகளுக்காக எவ்வாறு வரைமுறை படுத்தலாம் என்பதை மூன்றாம் இயலில் பார்த்தோம். சொற்பொருண்மை மயக்கநீக்கத்திற்கான பல நெறிமுறைளையும் வழிமுறைகளையும் வழிமுறை வரைவுகளையும் பற்றி நான்காம் இயலில் பார்த்தோம். இவ்வியலில் சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறை உருவாக்கத்திற்கு அடிப்படையாக அமையும் தமிழ்ச் சொற்பொருண்மையியல் ஆய்வும் இதுவரை தமிழில் சொற்பொருண்மை மயக்கநீக்கத்திற்காக மேற்கொள்ளப்பட்ட ஆய்வுகளின் சுருக்க உரையும் தரப்பட்டுள்ளது. மேலும் தமிழுக்கான சொற்பொண்மை மயக்கநீக்க ஒழுங்குமுறை ஒன்றை (ஒரு மாதிரியை) உருவாக்குவதற்கான முயற்சியும் மேற்கொள்ளப்படுள்ளது.

5.1 தமிழ்ச் சொற்பொருண்மையியல் ஆய்வு

சொற்களின் பொருள்கள் எல்லாம் சேர்ந்து வாக்கியத்தின் பொருளைத் தருவதாக வைத்துக் கொண்டால் சொற்களின் பொருளைப் பற்றி ஆய்வது இனியமையாதாகிறது. சொற்களின் பொருண்மையைப் பற்றி ஆய்வது சொற்பொருண்மையியல் (lexical semantics) ஆகும்.

5.1.1 சொற்களின் பொருண்மை உறவுகள்

சொற்களின் பொருண்மையை இருவகை உறவுகள் அடிப்படையில் அணுகலாம்:

1. உறுப்பமைவு உறவு அடிப்படையில் (by syntagmatic relation)

2. அடுக்கு உறவு அடிப்படையில் (by paradigmatic relation)

5.1.1.1 உறுப்பமைவு உறவு

ஒரு வாக்கியத்தில் ஒன்றையொன்று தொடர்ந்து வரும் சொற்களுக்கு இடையே உள்ள உறவு உறுப்பமைவு உறவாகும். பின்வரும் வாக்கியத்தில் *அவன், பாம்பை, கொன்றான்* என்ற சொற்களுக்கு இடையே உள்ள உறவு உறுப்பமைவு உறவாகும். *அவன்* என்பது எழுவாய் உறவையும் *பாம்பு* என்பது செயப்படுபொருள் உறவையும் *கொன்றான்* என்பது பயனிலை உறவையும் வெளிப்படுத்தி ஒன்றுக்கொன்று பொருண்மை உறவை வெளிப்படுத்தி நிற்கின்றன.

அவன்	பாம்பைக்	கொன்றான்
எழுவாய்	செயப்படுபொருள்	பயனிலை
செய்பவன்	அவதிப்படுவது	செயல்

5.1.1.2.அடுக்கு உறவு

ஒரு சொல்லுக்குப் பதிலாக அதே இடத்தில் வரும் சொற்களுக்கு இடையே உள்ள உறவு அடுக்குறவு எனப்படும். எடுத்துக்காட்டாகப் பின்வரும் வாக்கியத்தில் பாம்பு, பல்லி, பூனை, எலி என்பவைகளுக்கு இடையே உள்ள உறவு அடுக்குறவு எனப்படும்.

அவர் { பாம்பைக் / பல்லியைக் / பூனையைக் / எலியைக் } கொன்றான்

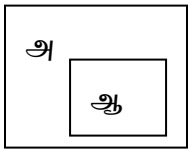
5.1..2 அடிப்படை உறவுகள்

வாக்கியங்களுக்கு இடையே உள்ள அடிப்படை உறவுகளை (Congruence Relations) நான்காகப் பகுத்து குருஸ் (Cruse, 1986) விளக்குகிறார். இவை சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளை நிறுவுவதற்கும் பிற அடுக்கு உறவுகளை விவரிப்பதற்கும் பயன்படும். அடிப்படை உறவுகளாவன ஒத்தல், உள்ளடங்கள், மேலுறல், வேறுபடல் (பிரிநிலை) என்பனவாகும். இவற்றைப் பின்வருமாறு வெளிப்படுத்தலாம்.

1. ஒத்தல் (identify)

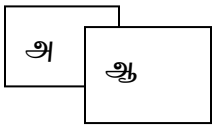
‘அ’ – ‘ஆ’ இவைகள் ஒரே அங்கங்களைப் பெற்றிருப்பவை.

2. உள்ளடங்கல் (inclusion)



‘அ’ வின் அங்கங்கள் ‘ஆ’ வின் அங்கங்களை உள்ளடக்கும்.

3. மேலுறல் (overlapping)



‘அ’ -விலும் ‘ஆ’ -விலும் ஒரே அங்கங்கள் காணப்படுவதோடு ஒன்றில் இல்லாத அங்கங்கள் மற்றொன்றில் காணப்படும்.

4. வேறுபடல் (பிரிநிலை) (disjunction)

‘அ’ –விலும் ‘ஆ’ –விலும் பொதுவான அங்கங்கள் காணப்படாது.

5.1.3. சொல்லுறவுகள்

லையான்ஸ் (1975) அடுக்குறவு முறையில் ஒன்றையொன்று மாற்றிக்கொள்ள இயலும் சொற்களுக்கிடையே உள்ள சொல்லுறவுகளைப் (lexical relations) பொருண்மை அடிப்படையில் பின்வரும் வகைப்படுத்தி ஆய்கிறார்.

- ஒருபொருள் பன்மொழியம் (Synonymy)
- உள்ளடங்கு மொழியம் (Hyponymy)
- இணக்கம் (Compatibility)
- இணக்கமின்மை (Incompatibility)

5.1.3.1 ஒருபொருள் பன்மொழியம்

பொருண்மை அடிப்படை ஒன்றையொன்று உள்ளடக்கும், இடம் பெயர்க்கும் சொற்களுக்கு இடையே உள்ள உறவு ஒருபொருள் பன்மொழியம் எனப்படும்.

புத்தகம்:நூல், உதிர்:விழு, காண்:பார், நுழை:பிரவேசி, விரட்டு: துரத்து.

ஒத்தருத்தலின் அளவுகோல்

ஒருபொருள் பன்மொழிய உறவை விரிவாக நோக்குதல் இன்றியமையாததாகிறது. இரண்டு நியாயமான பொருண்மை உள்ளுணர்வுகள் பின்வருவன:

1. சில இணைச் சொற்களும் சில கூட்டுச் சொற்களும் தம்மிடையில் ஒற்றுமையைக் காட்டுகின்றன. இந்த ஒற்றுமையை நாம் ஒருபொருள் பன்மொழியம் என்கிறோம்.
2. சில இணைச் சொற்கள் மற்ற இணைச் சொற்களை விடக் கூடுதலாக ஒத்திருக்கின்றன. நூல்:புத்தகம், தந்திரசாலி:புத்திசாலி

ஒருபொருள் பன்மொழிய பண்புகளைக் கூறச் சரியான வழியில்லை. கீழ்க்கண்ட இரு வழிகளை கூறலாம்.

1. முதலாவதாகத் தேவையான அளவு ஒற்றுமைகள், அனுமதிக்கக் கூடிய வேற்றுமைகள் மூலம்.
2. இரண்டாவதாகச் சூழல் அடிப்படையில் நிர்ணயிக்கப்பட்ட வாக்கிய சட்டத்தின் மூலம்.

ஒருபொருள் பன்மொழிகளுக்கு இடையில் தேவையான அளவு பொருண்மை மேலுறல் இருப்பது கட்டாயமாகும் என்பது வெளிப்படை. *தாழ்மையுள்ள, பணிவுள்ள* என்பவைகளுக்கிடையில் பொருண்மை மேலுறல் இருக்கின்றது. ஆனால் *உண்மையுள்ள, சிவப்பான* என்பவைகளுக்கிடையே பொருண்மை மேலுறல் இல்லை.

தாழ்மையுள்ள:பணிவுள்ள x உண்மையுள்ள:சிவப்பான

ஒருபொருள் பன்மொழிகள் மிக உயர்வான பொருண்மை மேலுறலைக் காட்டுவதுடன் மிகக் குறைவான வேற்றுமையைத் தான் காட்ட வேண்டும். பொதுவாக, இணை ஒருபொருள் பன்மொழிகளுக்கிடையில் ஒன்றை மறப்பது மற்றொன்றை மறுப்பதை ஒக்கும்.

நீ அந்த புத்தகத்தைப் படித்தாயா?

*இல்லை, அந்த நூலைப் படித்தேன்

அந்த புத்தகத்தைப் படித்தேன் = அந்த நூலைப் படித்தேன்

ஒருபொருள்பன்மொழிகள் முக்கியமான பொருண்மைப் பண்புகளில் ஒற்றுமையைக் காட்டும். வேறுபடுவதென்றால் முக்கியமில்லாத விளிம்பில் வரும் பண்புகளில் வேறுபடலாம். சிலவகை வெளிப்பாடுகளின் ஒருபொருள் பன்மொழிகள் சேர்ந்துவரலாம். எடுத்துக்காட்டாக, வரையறை விளக்கமாக அல்லது அர்த்தமாக ஒருபொருள்பன்மொழியைப் பயன்படுத்தப்படலாம்.

அவன் அவளைக் கொன்று விட்டான், அதாவது கொலை செய்து விட்டான்.

அவன் அவளை நேசிக்கிறான், அதாவது காதலிக்கிறான்.

சில சமயங்களில் ஒருபொருள் பன்மொழிகள் ஒன்றுக்கு ஒன்று முரண்படுவதுபோல் வரலாம். அப்பொழுது, 'அதாவது', 'சரியாகச் சொல்லப்போனால்' என்ற சொற்கள் இடையில் வரலாம்.

*அவன் ஒரு முட்டாள், சரியாகச் சொல்லப்போனால் ஒரு மடையன்.

பொதுவாக முரண்பட்டு வரும் சொற்கள் மேற்சொன்னப்டி வரா.

அவன் ஒரு முட்டாள், சரியாகச் சொல்லப்போனால் புத்திசாலி.

ஒருபொருள் பன்மொழிகளுக்குள் சில இணைச்சொற்கள் பிற இணைச் சொற்களைவிடக் கூடுதல் ஒப்புமையாக இருக்கும். இதனால் ஒப்புமையைக் காட்டுகிற அளவுகோல் உள்ளது போல் தோன்றும். இந்த அளவுகோலில் சரியான வரையறுக்கப்பட்ட இறுதி முனை தேவைப்படும். ஒருபொருள்பன்மொழியத்தின் நிரல் படி அந்த முனைதான் முழு ஒப்புமையைத் தெரியப்படுத்தும். புதிய ஒருபொருள் பன்மொழியை மறுமுனையில் பெறலாம்.

நீளம்……., கட்டை, பெரிது ……., சிறிது

ஒருபொருள் பன்மொழிகளையும் ஒருபொருள் பன்மொழிகள் அல்லாதவற்றையும் பிரிக்கும் கோடு தெளிவற்றது. பின்வரும் எடுத்துக்காட்டு இதனை வெளிப்படுத்தும்.

உலர்தல்:காய்தல், உணங்குதல், கரிதல்: கருகுதல்.

முற்று ஒருபொருள் பன்மொழியம்

இரண்டு சொற்கள் முற்று ஒருபொருள்பன்மொழியாக (absolute synonymy) அமைய வேண்டுமானால் அவற்றின் சூழல் உறவுகள் ஒன்றாக இருக்க வேண்டும். அத்தகைய முற்று ஒருபொருள் பன்மொழிகளைக் கண்டுபிடிப்பது நடைமுறைக்கு ஒவ்வாது. ஏனென்றால் அவை வரும் எல்லாச் சூழல்களையும் நோக்குவது இயலாத காரியம். ஆனால் அத்தகைய எதிர்பார்ப்பு நேரிடையாக முற்று ஒருபொருள் பன்மொழியை அறிய உதவும்.

முற்று ஒருபொருள்பன்மொழிகள் மிகக்குறைவாகவே உள்ளன. ஒரு மொழியில் முற்று ஒருபொருள் பன்மொழியின் தேவைக்குச் சரியான காரணமில்லை. அப்படியிருக்குமானால் போட்டியிடும் இணைச் சொற்களில் ஒன்று இல்லாமல் போகவோ அல்லது அவற்றின் பொருண்மைச் செயற்பாடு மாறவோ செய்யும். பின்வரும் எடுத்துக்காட்டுகள் இதை விளக்கும்.

நூல்:புத்தகம், சொல்லுதல்:கூறுதல்

புலனறி ஒருபொருள் பன்மொழியம்

இரு சொற்கள் புலனறிவு ஒருபொருள்பன்மொழிகளாக (cognitive synonymy) இருக்க வேண்டுமானால் அவற்றின் பொருண்மைப் பண்புகளில் சில ஒன்றாய் இருக்கவேண்டும். புலனறிவு ஒருபொருள்பன்மொழிகளுள் சிலதான் முற்று ஒருபொருள்பன்மொழிகளாக அமைகிறது. இதிலிருந்து அறிவது என்னவென்றால் பெரும்பான்மையான இணைகள் சிலவகையில் பொருள் வேறுபட்டு தான் இருக்கின்றன.

பொருள்கொள்ளும் விதம்

கீழ்க்கண்டவை பொருள்கொள்ளும் (semantic mode) விதத்தில் வேறுபடுகின்றன.

அ. எனக்கு வலிக்கிறது

ஆ. ஐயோ!

முதல் வாக்கியத்தில் பொருள்கொள்வது கூற்று விதமாகும் (propositional mode). இரண்டாம் வாக்கியத்தில் பொருள்கொள்வது வெளிப்படும் விதமாகும் (expressive mode); இரண்டும் வெவ்வேறு ஆகும். கூற்றுப் பொருண்மைக்கும் வெளிப்பாட்டுப் பொருண்மைக்கும் வேற்றுமைகள் உள்ளன. வெளிப்பாட்டுப் பொருண்மை உண்மை நிலையை (truth-condition) நிச்சயிப்பதில் எந்தவித பங்கும் வகிப்பதில்லை. வெளிப்படை "அது பொய் உனக்கு வலிக்கவில்லை" என்று மறுக்க இயலாது.

வெளிப்பாட்டுப் பொருண்மை மன உணர்ச்சியையோ மனோபாவத்தையோ தெரிவிப்பதாக அமையும். சந்தேகம், நிச்சயம், எதிர்பார்ப்பு, ஆச்சரியம், வெறுப்பு, ஏமாற்றம் போன்றவை அடங்கும். பின்வரும் வாக்கியங்கள் இதைத் தெளிவுபடுத்தும்.

அவன் இதுவரை வரவில்லையா?

அவன் ஏற்கனவே வந்துவிட்டானா?

அவன் இன்னும் வரவில்லையா?

வெளிப்பாட்டுப் பண்பும் கூற்றுப்பண்பும் ஒரு சொல்லின் பொருண்மையில் விரவி இருக்கலாம். எடுத்துக்காட்டாக, *அப்பா*, *அம்மா* என்பவனவற்றில் வெளிப்பாட்டுப் பண்பும் கூற்றுப்பண்பும் விரவிவரும். வெளிப்பாட்டுப் பொருள் மிகவும் முக்கியமாக அமையலாம். வெளிப்பாட்டுப் பொருண்மை இல்லாமல் புலப்படுத்தம் (communication) அமைவது முடியாது என்று கூட வாதிடலாம். சில சொற்கள் பிற சொற்களைக் காட்டிலும் கூடுதல் வெளிப்பாட்டை உணர்த்துவதாய் அமையலாம். *அம்மா, அப்பா* என்பன கூடுதல் வெளிப்பாட்டையும், *தாய், தந்தை* என்பன குறைந்த வெளிப்பாட்டையும் உணர்த்தும்.

அம்மா:தாய், அப்பா:தந்தை

வெளிப்பாட்டுப் பண்புகளால் மட்டும் வேறுபடும் சொற்கள் எல்லாம் புலனறிவு ஒருபொருள் பன்மொழிகளாகும்.

முன்னதாகப் பெறும் பொருண்மையும் தூண்டப்படும் பொருண்மையும்

முன்னதாகப் பெறும் பொருண்மைக்கும் (pre-supposed meaning) தூண்டப்படும் பொருண்மைக்கும் (evoked meaning) வேறுபாடு பாராட்டலாம். "குடித்தல்" குடிப்பவரை எழுவாயாய் எதிர்பார்க்கிறது. அதுபோல் "இறத்தல்" ஒரு விலங்கை எழுவாயாக எதிர்பார்க்கிறது. இம்மாதிரிப்பட்ட உடன் வருகை (co-occurance) கட்டுத்திட்டத்தைச் சேர்ந்துவருகைக் கட்டுத்திட்டம் (collecational condition) எனலாம். பின்வரும் எடுத்துக்காட்டு இதைத் தெளிவுபடுத்தும்.

அவர் இறைவனடி எய்தினார்.

அவர் துஞ்சினார்.

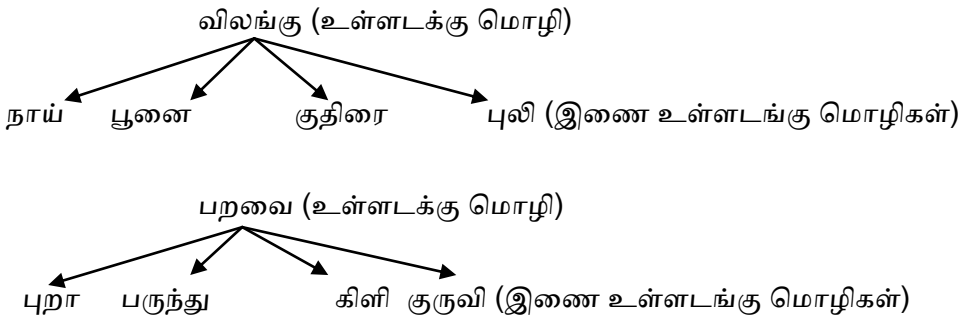
அவர் இறந்தார்.

*நாய் துஞ்சினது.

சேர்ந்துவருகைக் (collecational) கட்டுத்திட்டங்கள் அளபு அடிப்படையில் வேறுபடலாம். சிலவற்றை ஒழுங்கான சேர்ந்துவருகைக் கட்டுத்திட்டம் என்றும் சிலவற்றை பகுதி ஒழுங்கான சேர்ந்துவருகைக் கட்டுத்திட்டம் என்றும் கூறலாம்.

5.1.3.2. உள்ளடங்கு மொழியம் (hyponymy)

பொருண்மை அடிப்படையில் உள்ளடங்கும் சொற்களுக்கும் உள்ளடக்கும் சொற்களுக்கும் இடையே உள்ள உறவு உள்ளடங்கு மொழியம் எனப்படும்.



எதிர்மை (opposition), முரண் (contrast) ஆகியவற்றைப் போல உள்ளடங்கு மொழியமும் மிக முக்கியமான அடுக்கு உறவாகும். முரண் தனிப்பட்ட அல்லது கீழ் அடங்கும் சொல்லுக்கும் பொது அல்லது உள்ளடக்கும் சொல்லுக்கும் உள்ள உறவைக் குறிப்பதாம்.

பசு:விலங்கு, ரோஜா:பூ, வாங்கு:பெறு

ரோஜா, முல்லை, மல்லிகை:பூ, ஆடு, பசு, காளை:விலங்கு

தர்க்கவியலார் (logicinans) வகுப்பில் அடக்கல் (class inclusion) என்ற நிலையில் உள்ளடங்கு மொழியத்தை விவரிக்கிறார்கள்.

அ ⊃ ஆ ஆ ⊂ அ

பூ ⊃ ரோஜா ரோஜா ⊂ பூ

எல்லா ரோஜாக்களும் பூக்கள், ஆனால் பூக்கள் எல்லாம் ரோஜாக்கள் அல்ல. தர்க்கவியலார் விளக்கத்தில் தெளிவின்மை இருக்கிறது. உள்ளடங்குமொழி உள்ளடக்கு மொழியில் அடங்குகிறதா உள்ளடக்கு மொழி உள்ளடங்கு மொழியில் அடங்குகிறதா என்ற ஐயம் எழுகிறது. சொற்களின் அகலத்தைப் (extension) பார்த்தால் உள்ளடங்குமொழி கூடுதல் உள்ளடக்கும் தன்மையுடையது என்றும் சொற்களின் ஆழத்தைப் (intension) பார்த்தால் உள்ளடங்கு மொழி கூடுதல் உள்ளடக்கும் தன்மையது என்றும் கூறலாம். எடுத்துக்காட்டாக, ரோஜா, பூவின் எல்லாப் பண்புகளையும் பெற்றிருப்பதோடு கூடுதலாக ரோஜாவின் தனித்தன்மையான பண்புகளையும் பெற்றிருக்கிறது.

உள்ளடங்கு மொழியத்தை ஒருதலை நோக்கு உறவு (unilateral relation) மூலம் வரையறை விளக்கம் செய்யலாம். *அவள் தலையில் மல்லிகை வைத்திருக்கிறாள்* என்பது *அவள் தலையில் பூ*

வைத்திருக்கிறாள் என்பதை உள்ளடக்கும். ஆனால் *அவள் தலையில் பூ வைத்திருக்கிறாள்* என்பது *அவள் மல்லிகை வைத்திருக்கிறாள்* என்பதை உள்ளடக்காது. இதன்படி ஒருபொருள் பன்மொழியத்தை இரு நோக்கு உறவு (bilateral relation) மூலம் ஒத்த தன்மையை (symmetry) விவரிக்கலாம். உள்ளடங்கு மொழியத்தை ஒவ்வாத் தன்மையாகக் கருதலாம் (asymmetry).

உள்ளடங்கு மொழியத்தை ஒரு கடவு (transitive) உறவாகக் கொள்ளலாம். 'அ' என்பது 'இ' யின் உள்ளடங்கு மொழியாகவும், 'இ' என்பது 'உ' வின் உள்ளடங்கு மொழியாகவும் வருமானால், 'அ' என்பது 'உ'வின் உள்ளடங்கு மொழியாகும்.

எ.கா.

பசு⊂பாலூட்டி, பாலூட்டி⊂விலங்கு என்றால் பசு⊂விலங்கு ஆகும்.

உள்ளடங்கு மொழியை கீழ்வரும் சூத்திரத்தில் ஏற்றி பகுப்பாய்வு (analytic) நியாயத்தை வெளிப்படுத்த முடியும்.

அ என்பது ஒரு வகை இ ('அ', 'இ' என்பதன் உள்ளடங்குமொழி என்றால்)

பசு என்பது ஒரு வகை விலங்கு / பசு ஒரு விலங்கு.

'இ' என்ற சொல் ஒன்றுக்கும் மேற்பட்ட உள்ளடங்குச் சொற்களை உள்ளடக்கும் என்றால் கீழ்வரும் வாக்கியங்களில் 1, 2 என்பன சரியானதாகும்.

பசுக்களும் வேறு விலங்குகளும் வந்தன.

ரோஜாவும் வேறு பூக்களும் வாங்கினாள்.

? பசுக்களும் வேறு பூக்களும்.

? ரோஜாவும் வேறு விலங்குகளும்.

இம்மாதிரிப்பட்ட வெளிப்பாடுகள் (வாக்கிய வெளிப்பாடுகள்) உள்ளடங்கு மொழியத்தையும் உள்ளடக்கு மொழியத்தையும் நிறுவுவதற்கு முக்கியமாகும். இதில் முக்கியமாகக் கவனிக்க வேண்டியது என்னவென்றால் சொற்களின் பொருளை அறியாமலேயே ஒரு சொல் மற்றொரு சொல்லின் உள்ளடங்கு மொழி என்றும் ஒரு சொல்லும் மற்றொரு சொல்லும் இணை உள்ளடங்கு மொழிகள் என்றும் அறியலாம்.

குவளை ஒருவகை மலர். / குவளை ஒரு மலர்.

வங்கு ஒருவகை விலங்கு. / வங்கு ஒரு விலங்கு.

நம் மொழியிலுள்ள சொற்களின் பொருளைப் பற்றி நமது அறிவு பெரும்பாலும் மேற்சொன்ன விதமானதாகும். எடுத்துக்காட்டாக, நாம் "ஆல்" என்பது "மரம்" என்ற சொல்லின் உள்ளடங்கு மொழியாகும் அல்லது "மீன்கொத்தி" ஒரு "பறவை" ஆகும் என்று அறிவோமேயன்றி ஆல் பிற மரங்களிலிருந்து எவ்வாறு வேறுபடுகிறதென்றோ மீன்கொத்தி எவ்வாறு பிற பறவைகளிலிருந்து வேறுபடுகிறது என்றோ சரியாகக் கூறுவதற்கு அறியோம்.

பெரும்பாலும் உள்ளடங்கு மொழி ஒரு பெயரடையை உள்ளடக்கி இருப்பதால் அவ்வடையும் உள்ளடக்கு மொழியும் சேர்ந்த தொடர் உள்ளடங்கு மொழிக்கு இணையாய் வருவதுண்டு.

ஒரு யானையைப் பார்த்தேன்

ஒரு பெரிய விலங்கைப் பார்த்தேன்

இதனால் ஒரு உள்ளடங்கு மொழியும் அதனோடு தொடர்புடைய பெயரடையும் உள்ளடக்கு மொழியும் சேர்ந்த தொடரும் ஒருபொருள்பன்மொழிகள் என்று அர்த்தமல்ல. எடுத்துக்காட்டாக, *பசு* என்பதும் *கொம்புள்ள விலங்கு* என்பதும் ஒருபொருள் பன்மொழிகள் அல்ல. உள்ளடங்கு மொழியும் அதனை உள்ளடக்கும் உள்ளடக்குமொழியும் சேர்ந்து வருவதுண்டு.

முல்லைப்பூ

ஆல் மரம்

இதுபோல் (மேற்சொன்னது போல்) வரும் கூட்டுச் சொற்கள் சில சமயம் நமக்குத் தவறான எண்ணத்தை உருவாக்கலாம் எடுத்துக்காட்டாக, *கழுதைப்புலி* என்பது ஒருவகைப் புலியல்ல; அதுபோல் *வரிக்குதிரை* என்பது ஒரு வகைக் குதிரையல்ல.

வினைகளையோ பெயரடைகளையோ வினையடைகளையோ பிற சொல் வகைகளையோ 'அ' என்பது ஒரு வகை 'ஆ' என்ற சூத்திரத்தில் அடக்கிவிட இயலாது. எல்லாச் சொல் வகைகளுக்கும் உள்ளடங்கு மொழியத்தை நிரூபிப்பது என்பது கடினமான காரியம். பொதுவாகச் சொல்லப்போனால் உள்ளடக்கு மொழியும் தொடர்புடைய உறுப்பமை அடையும் சேர்ந்த கூட்டும் உள்ளடக்கு மொழியும் அடுக்குறவில் வருவது தான் உள்ளடங்குமொழியம்.

5.1.3.3 இணக்கம்

பொருண்மை அடிப்படையில் ஒன்றையொன்று மேலுறலாக வரும் சொற்களுக்கு இடையே உள்ள உறவு இணக்கம் எனப்படும். இணங்கிவரும் சொற்கள் சில பொருண்மைப் பண்புகளில் ஒன்றுபட்டும் சில பொருண்மைப் பண்புகளில் வேறுபட்டும் இருக்கும்.

பாம்பு: விஷஜந்து (சரியான இணக்கம் (strict))

நாய்: செல்லப் பிராணி (தற்செயலான இணக்கம் (contigent))

5.1.3.4. இணக்கமின்மை

பொருண்மை அடிப்படையில் உள்ளடங்கவோ உள்ளடக்கவோ செய்யாமல் இணக்கமின்றி வேறுபட்டு வரும் சொற்களுக்கு இடையே உள்ள உறவு இணக்கமின்மை (incompatibility) எனப்படும்.

மலை : மாடு

ஆடுதல் : தின்னுதல்

5.1.3.5. எதிர்மொழியம்

இணக்கமற்ற ஆனால் ஒன்றையொன்று எதிர்ப்பதுபோல் பொருள் அமையும் சொற்களுக்கு இடையே உள்ள உறவு எதிர்மொழியம் (antonymy) எனப்படும்.

எ.கா

உயரம்	:	குள்ளம்
நல்லது	:	கெட்டது
நீளம்	:	குட்டை
விரைவாக	:	மெதுவாக
இலகு	:	கடினம்

எதிர்மொழியத்தின் பண்புகள்

1. எதிர்மொழிகளைக் கிரமப்படுத்தலாம் (gradable)
2. இணையாக வரும் எதிர்மொழிகள் நீளம், வேகம், கனம் போன்ற பண்புகளை உணர்த்தும்
3. அப்பண்புகளைத் தீவிரமாக்கினால் இணை எதிர்மொழிகள் குறிப்பிட்ட பண்பு காட்டுகிற அளவுகோலில் எதிர்நோக்கிச் செல்லும்.

   ரொம்ப கனம் > கனம் > இலேசு > ரொம்ப லேசு

4. இணை எதிர்மொழிகள் ஒரு பொருள் பரப்பை இரண்டாகப் பிரிக்காது. அதாவது எதிர்மொழிகள் காட்டுகிற பொருண்மைகளுக்கு இடையில் படிப்படியாக வேறுபடுகிற பண்பு இருக்கும். இதனால் எதிர்மொழிகளைக் கொண்டு இரண்டு வேறுபட்ட ஆனால் ஒன்றுக்கு ஒன்று முரண்படாத வாக்கியங்களாக மொழியலாம்.

அது நீளமானது: அது குட்டையானது

இவ்வாக்கியங்கள் ஒன்றுக்கொன்று வேறுபட்டதேயன்றி முரண்பட்டதல்ல மற்றும் கீழ்வரும் வாக்கியமும் தப்பானது அல்ல.

அது நீளமும் அல்ல குட்டையும் அல்ல

எதிர்மொழியத்தின் உட்பிரிவுகள்

எதிர்மொழியங்களை மூன்றாகப் பிரிக்கலாம்.

1. துருவ எதிர்மொழியம் (polar antonyms)
2. மேலுறல் எதிர்மொழியம் (overlapping antonyms)
3. சமசக்தியுள்ள எதிர்மொழியம் (equipollent antonyms)

துருவ எதிர்மொழியம் (polar antonyms)

எடுத்துக்காட்டாகப் பின்வரும் இணைகள் துருவ எதிர்மொழியத்தின் கண் படும்.

கனம்:இலேசு, விரைவாக:மெதுவாக, உயரம்:குள்ளம், அகலம்:குறுக்கம், கட்டி:மெலிது, சிரமம்:இலகு

துருவ எதிர்மொழியம் பின்வருமாறு வரும்.

அவன் குள்ளமானவன் ஆனால் அவளைவிட உயரமானவன்.

அவன் உயரமானவன், ஆனால் அவளைவிட குள்ளமானவன்.

மேலுறல் எதிர்மொழியம் (overlapping antonyms)

மண்டு:புத்திசாலி, நல்லது:கெட்டது, இரக்கம்:கொடூரம்

அவன் மண்டு, ஆனால் அவளைவிட புத்திசாலி

அவன் புத்திசாலி, ஆனால் அவளைவிட மண்டு

துணைநிலை (complementarily)

துணைநிலைச் சொற்கள் ஒரு கருத்துருப்பரப்பை (conceptual domain) ஒன்றுக்கொன்று தனிப்பட்ட இரண்டு பகுதிகளாகப் பிரிக்கும்.

உண்மை:பொய், திற:மூடு, வெற்றி:தோல்வி

இது இணை துணைச் சொற்களில் ஒன்று காட்டும் பொருளை ஏற்பது மற்றதன் பொருளை மறுப்பதை ஒக்கும். *கதவு திறந்திருக்கிறது* என்றால் *கதவு மூடியிருக்கவில்லை* என்று அர்த்தம். *கதவு திறந்திருக்கவும் இல்லை, மூடியிருக்கவும் இல்லை* என்பது சரியில்லை. *அவள் தேர்வில் வெற்றியடைந்தான்* என்றால் *அவள் தேர்வில் தோல்வியடையவில்லை* என்று அர்த்தம். *அவள் தேர்வில் வெற்றியடைவுமில்லை தோல்வியடையவுமில்லை* என்பது சரியல்ல.

மறுதலை (conversances)

இரண்டு சொற்களில் ஒன்றின் பொருள் மற்றதின் மறுதலையாக வருமானால் அவைகளுக்கு இடையே உள்ள உறவை மறுதலை என்கிறோம்.

கணவன்:மனைவி, குழந்தை:பெற்றோர்

இராஜா ராணியின் கணவன் என்றால் *ராணி இராஜாவின் மனைவி* என்று அர்த்தம். இராதை அவர்கள் குழந்தை என்றால் *அவர்கள் இராதையின் பெற்றோர்கள்* என்று அர்த்தம்.

பரஸ்பர சமூகப் பங்களிப்பளர்கள்

ஒரு சமூகத்தில் பங்களிப்பு செய்யும் பங்களிப்பாளர்களுக்கு இடையில் ஒருவித தொடர்பு அல்லது பரஸ்பர சார்பு காணப்படும். பின்வரும் இணைகள் எடுத்துக்காடுகளாகும்.

மருத்துவர்:நோயாளி, முதலாளி:தொழிலாளி, எஜமானி:வேலைக்காரி

உறவுமுறைகள்

உறவு முறையை வெளிப்படுத்தும் சொற்களுக்கிடையில் 'உடைமை' உறவு காணப்படும்.

அப்பா/அம்மா: மகன்/மகள்

இராஜா இராதையின் அப்பா என்றால், *ராதை இரஜாவின் மகள்* ஆவாள் என்று பொருள்.

கால இடைநிலை உறவுகள்

முன்னே:பின்னே, முன்னால்:பின்னால், மேலே:கீழே

திசை எதிர்மறைகள் (Directional oppositions)

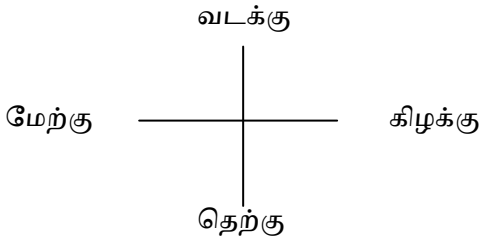
வா	:	போ
மேலேறு	:	கீழிறங்கு
முன்னேறு	:	பின்னேறு
சேர்	:	புறப்படு

வலது : இடது

முன் : பின்

செங்கோண எதிர்நிலை உறவும் துருவ எதிர்நிலை உறவும் (orthogonal and antipodal opposition)

இவ்வகையில் படும் எதிர்நிலைகளைத் திசைச் சொற்களால் விளக்கலாம்.



செங்கோண எதிர்நிலை உறவு: இவ்வுறவு காட்டும் இணைகளில் ஒவ்வொரு சொல்லும் பிற இரண்டு சொற்களுடன் செங்கோண எதிர்நிலையில் வரும்.

வடக்கு:கிழக்கு மற்றும் மேற்கு, கிழக்கு:தெற்கு மற்றும் வடக்கு

சிறுமி:சிறுவன் மற்றும் பெண், girl: boy மற்றும் woman

துருவ எதிர்நிலை உறவு: இவ்வுறவு காட்டும் இணைகளில் ஒவ்வொரு சொல்லும் மற்ற சொல்லுடன் துருவ எதிர்நிலையில் உள்ளது.

வடக்கு:தெற்கு, கிழக்கு:மேற்கு

மேலே:கீழே, முன்னால்:பின்னால், இடது:வலது

துருவ எதிர்சொற்கள் இடச்சொற்களை மட்டுமின்றி பிற சொற்களையும் உள்ளடக்குகிறது.

எ.கா

நிறங்கள் – கறுப்பு: வெள்ளை; சிவப்பு: கறுப்பு

எதிர்மொழி என்பது 'ஒருபோல்' என்பதிலிருந்து மிக வேற்பட்டு எதிர்ப்புறமாக இருப்பது என்று கருதப்பட்டது. அது சரியல்ல. நாம் இரண்டு பொருள்களை ஒப்பிடவோ வேறுபடுத்தவோ செய்கையில் நாம் அவைகளுக்குள்ள வேறுபாட்டையும் ஒற்றுமையும் பார்த்து என்ன குணங்களில் வேறுபடுகின்றன என்று கூறுகிறோம்.

எ.கா.

married : single

ஒற்றுமைகளை நோக்கித்தான் எதிர்மொழிய உறவு நிறுவப்படுகிறது.

5.1.3.6 பகுதி-முழுமை உறவுகள் (part and whole relations/Meronymy-Holonymy relations)

பகுதி-முழுமை உள்ளடங்கு உறவிலிருந்து மாறுபட்டாகும். எடுத்துக்காட்டாகப் பின்வரும் இணைகள் பகுதி-முழுமை உறவை வெளிப்படுத்தும்: *கை:உடம்பு, சக்கரம்:வண்டி. கை* ஒருவகை உடம்பு அல்ல. உடம்பின் பாகம்தான். உடம்பின் ஒரு பகுதி *கை* ஆகும். பகுதி-முழுமை உறவு உள்ளடங்கு உறவைப் போல கடவு (transitive) உறவாகும். 'அ' என்பது 'ஆ' என்பதன் பாகம் என்றும் 'ஆ' என்பது 'இ' வின் பாகம் என்று கொண்டால் 'அ' என்பது 'இ' என்பதன் பாகமாகும். எடுத்துக்காட்டாக, *கண்* என்பது *தலை* என்பதன் பாகம்: தலை என்பது உடம்பு என்பதன் பாகம்; எனவே கண் என்பது உடம்பு என்பதன் பாகமாகும். முழுமையைக் குறிபிடும் சொற்கள் முழுமொழிகள் (holonyms) என்றும் சினையைக் குறிப்பிடும் சொற்கள் சினைமொழிகள் (meronyms) என்றும் அழைக்கப்படும்.

ஒரு சொல் மற்றொரு சொல்லின் பாகம் என்ற காரணத்தால் அவைகளுக்குள் பகுதி-முழுமை உறவு இருப்பதாகக் கூற இயலாது. எடுத்துக்காட்டாக *கைப்பிடி* என்பது *கதவு* என்பதன் பாகம், *கதவு* என்பது *வீடு* என்பதன் பாகம் என்ற காரணத்தால் கைப்பிடி வீட்டின் பாகம் என்று கூறுவது கடினம். "கதவுக்கு கைப்பிடி இல்லை" என்று கூறலாம். ஆனால் வீட்டிற்கு *கைப்பிடி* இல்லை என்று கூற முடியாது.

எனவே, கதவுக்கும் வீட்டுக்கும் உள்ள உறவு பகுதி-முழுமை உறவாகும். அதுபோல் கைப்பிடிக்கும் கதவுக்கும் உள்ள உறவு பகுதி-முழுமை உறவாகும். ஆனால் கைப்பிடிக்கும் கதவுக்கும் உள்ள உறவு பகுதி-முழுமை உறவாகாது. கதவும் வீடும் வெவ்வேறு வகையான முழுமைகளாகும். ஒரு மொழியின் சொற்றொகுதியின் பலசொற்களின் பொருண்மைகளை 'பகுதி-முழுமை' உறவைக் கூறாமல் வரையறை விளக்கம் செய்ய இயலாது. பின்வரும் இணைகளில் வரும் பகுதிகளை முழுமையைக் கூறாமலோ முழுமையைப் பகுதியைக் கூறாமலே விளக்கவியலாது.

வினாடி:நிமிடம்:மணி:நாள்:வாரம்:மாதம்:ஆண்டு

சொற்றொகுதியின் உள்ளடக்குச் சொல்லாக வரும் தொகைப் பெயர்கள் (collective nouns) படிநிலை உறவில் உள்ளடங்கு மொழிய சொற்கள் போலவும் 'பகுதி-முழுமை' சொற்கள் போலவும் வரும்.

கால்நடைகள்: பசு, காளை, ……..

மரத்தால் செய்யப்பட்ட பொருட்கள்: நாற்காலி, மேஜை .....

பசு, காளை என்பனவற்றை கால்நடைகளில் ஒருவகையாகவும் கருதலாம். உறுப்பினர்களாகவும் கருதலாம். அதுபோல் *நாற்காலி, மேஜை* என்பனவற்றை மரத்தால் செய்யப்பட்ட பொருட்களின் வகைகளாகவும் கொள்ளலாம். உறுப்பினர்களாகவும் கொள்ளலாம். தொகைப் பெயர்கள் பொருண்மை அடிப்படையில் திரப்பெயர் (mass noun) போல இருப்பதால் இவ்வாறு இரு உறவு வர இயலுகிறது.

வேறுவகையான தொகைப்பெயர்களுக்கு எடுத்துக்காட்டு *மந்தை, நூலகம், காடு* என்பன. ஆட்டுக்கும் மந்தைக்கும் உள்ள உறவு உள்ளடங்கு உறவல்ல. *ஆடு ஒரு வகை மந்தை* என்று கூறவியலாது. அவ்வுறவை "*கை : உடம்பு*" போன்ற 'பகுதி-முழுமை' உறவு என்று கூறவியலாது. மந்தையில் 'ஆடு' என்ற ஒரே மாதிரியான உறுப்பினர்கள் தான் உள்ளன. ஆனால் உடம்பில் பலமாதிரியான உறுப்புகள் உள்ளன.

பாகத்தைத் துண்டுகளிலிருந்து வேறுபடுத்திக் காட்ட வேண்டும். 'பகுதி-முழுமை' உறவு 'துண்டுகள்' உறவிலிருந்து வேறுபட்டவை. பாகத்தின் மூன்று முக்கியமான பகுதிகள்: தன்னியக்கம், இடுகுறியற்ற எல்லை, குறிப்பிடத்தகுந்த செயற்பாடு.

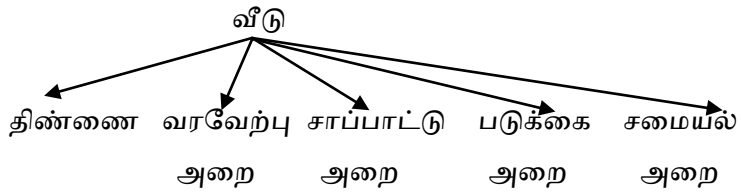
1. தன்னியக்கம்: சக்கரம் சைக்கிளின் ஒரு பாகம்.
2. இடுகுறி அற்ற எல்லை: பாகங்களாகப் பிரிக்கப்படும் எல்லைகள் வரையறுக்கப்பட்டவை. எடுத்துக்காட்டாக *மேற்கை, உள்ளங்கை, முன்னங்கை* போன்றவை மூட்டுகளால் இணைக்கப்பட்டவை.
3. குறிப்பிட தகுந்த செயல்பாடு: பாகங்களுக்குக் குறிப்பிடத்தகுந்து செயல்பாடு இருக்கும். எடுத்துக்காட்டாக, கண் பார்ப்பதற்கும், பிரேக் நிறுத்துவதற்கும் பயன்படுகிறது.

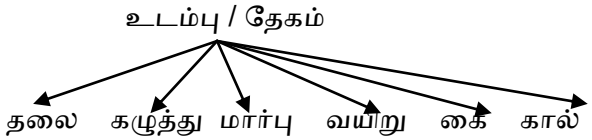
'பகுதி-முழுமை' உறவை ஆங்கிலத்தில் part and whole relations என்று கூறுவார்கள்.

தலை, கால் .....segmental parts,

எலும்பு, தசை, நாளம் .....systemic parts

பல முழுப்பெயர்களை உடம்பு போல் பல பாகங்களாகப் பிரிக்கலாம். எடுத்துக்காட்டாக வீடும் வீட்டின் பாகங்களையும் கூறலாம்.





பகுதி-முழுமை போன்ற உறவுகள்

பகுதி-முழுமை போன்ற பல்வேறு உறவுகள் சொற்களுக்கிடையில் வருவதைக் காணலாம். இவ்வேறுபாடுகளைப் பல பரிணாமணங்களில் உணரலாம்.

1. இடம் – இடத்தின் ஒருபகுதி

   தாலுகா: மாவட்டம்: இராஜியம்: தேசம்

   தலைநகர்: தேசம்

2. குழு – அங்கத்தினர் உறவு (group-member relation)

   குடும்பம்: குடும்பத்தினர்

3. வகுப்பு – அங்கத்தினர் உறவு (class-member relation)

   குவியல்: நெல்

   காடு: மரம்

   நூலகம்: நூல்

4. உறுப்புகள் அல்லது கூட்டுப்பொருளின் ஒரு பகுதி (constituents or ingredients)

   மசாலா: மிளகு, ஜீரகம்

5. பொருள் : பருப்பொருள் (obiect : material)

   ஆடை: நூல்

6. சாதனம் : துகள் (substance : particle)

   மழை: துளி

5.1.4. பிற பொருண்மை உறவுகள்

5.1.4.1. பகுதி உறவுகள்

தொடரியல் வருகை முறையில் பகுதி மேலுறலாக வரும் சொற்களுக்கு இடையே உள்ள உறவு பகுதி உறவு (partial relation) எனப்படும். பின்வரும் எடுத்துக்காட்டுகள் இதனைத் தெளிவுப்படுத்தும்.

ஒளி: மறை

அவன் தன் பணத்தை ஒளித்து வைத்தான்.

அவன் தன் பணத்தை மறைத்து வைத்தான்.

அவன் அவள் கண்களை மறைத்தான்.

*அவன் அவள் கண்களை ஒளித்தான்.

சூரியன் மறைந்தது.

*சூரியன் ஒளித்தது.

5.1.4.2. முழுமையுறா உறவுகள்

சில சமயங்களில் சரியான இணை இல்லாததால் சொற்களுக்கு இடையே உள்ள அடுக்கு உறவு முழுமையுறாது நிற்கும். வேண்டிய பொருளைக் கொண்டு ஆனால் வேறு தொடரியல் வகையைச் சார்ந்த சொல் அந்த இடைவெளியைப் பூர்த்தி செய்யுமாறு அமையும். இத்தகைய நிலையில் உள்ள உறவை முழுமையுறா உறவு (quasi-relation) எனலாம். எடுத்துக்காட்டாக, *நிறம்* என்ற பெயருக்கும் *செம், கரும், பசும், வெண்* என்ற பெயரடைகளுக்கும் இடையே உள்ள உறவு முழுமையுறா உறவாகும்.

நிறம்: செம், கரும், பசும், வெண்

5.1.4.3. போலி உறவுகள்

சொற்களுக்கு இடையே குறிப்பிடத்தக்க உறவு இல்லாமல் ஆனால் சூழ்நிலை மூலம் உறவு தோன்றினால் அச்சொற்களுக்கு இடையே உள்ள உறவைப் போலி உறவு எனலாம். உண்மையான ஒருபொருள் பன்மொழிகள் போலி ஒருபொருள் பன்மொழிகளிலிருந்து வேறுபடும். பின்வரும் எடுத்துக்காட்டுகளில் *கோணம்* என்பதும் *பக்கம்* என்பதும் போலி ஒருபொருள் பன்மொழிகள் ஆகும்.

முக்கோணத்திற்கு மூன்று கோணங்கள் உண்டு

முக்கோணத்திற்கு மூன்று பக்கங்கள் உண்டு

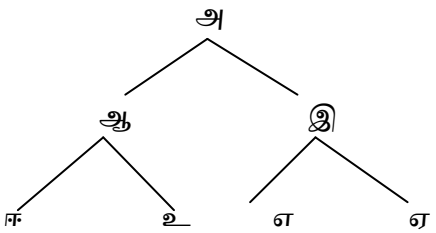
கோணம்:பக்கம் (போலி ஒருபொருள்பன்மொழிகள்)

5.1.5. சொற்றொகுதியின் பொருண்மை அமைப்பு

ஒரு மொழியின் சொற்றொகையைப் படிநிலை அமைப்பாகக் (hierarchical structure) காணவியலும். சொற்கள் ஒன்றுக்கொன்று தொடர்பு கொண்டிருப்பதைப் படிநிலையாகவும் (hierarchy) விகிதத்தொடர்களாகவும் (propotional series) காணலாம்.

5.1.5.1. படிநிலை அமைப்பு

இரண்டு வகையான படிநிலை அமைப்பைக் (hierarchial structure) காணலாம். ஒன்று கிளைகளாலானவை (branching) மற்றொன்று கிளைகற்றவை (non-branching). இவற்றைக் கீழ்க்காணும் படங்கள் மூலம் வேறுபடுத்திக் காட்டலாம்.



‘ஆ’ வுக்கும் ‘அ’ வுக்கும் உள்ள உறவு செங்குத்தான உறவு (vertical relation) உயர் உறவாகும் (relation of dominance). ‘ஆ’ வுக்கும் ‘இ’ வுக்கும் உள்ள கிடைநிலை உறவு (horizontal relation) வேறுபடு உறவாகும் (relation of difference).

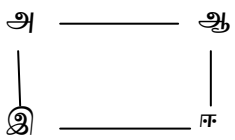
உயர் உறவு ஒவ்வாதிருக்கும் (asymmetric). அது ஒருவழி திசை நோக்குப் பண்புகளைப் பெற்றிருக்கும். ஒத்திருக்கும் உறவு இருவழி திசைநோக்குப் பண்பைப் பெற்றிருக்கும். உயர் உறவு தொடர்ச்சியாகச் சங்கிலிபோல் அமையலாம்.

ங ஞ ண ந ம ன

உயர் உறவு கடவு (transitive) உறவாக இருக்கலாம் அல்லது கடவல்லா (intransitive) உறவாக இருக்கலாம். அ⊃ஆ, இ⊃ஈ என்றால் அ⊃ஈ என்று அர்த்தம் வருவது கடவு உறவாகும். அ⊃ஆ, இ⊃ஈ ஆனால் அ⊅இ என்று அர்த்தம் தருவது கடவல்லா உறவாகும்.

5.1.5.2. விகிதத் தொடர்கள்

எளிமையான விகிதத்தொடர்கள் (propotional series) நான்கு தனிமங்கள் அடங்கிய ஒரு சட்டத்தைக் கொண்டிருக்கும்.



ஏதாவது மூன்று தனிமங்களிலிருந்து நான்காவதன் உறவு தீர்மானிக்கப்படுமாறு தனிமங்களுக்கிடையில் உறவு இருக்க வேண்டும். பின்வரும் விகித உறவுகள் வருமாறு சட்டம் அமைக்கப்படும்.

அ–வுக்கும் ஆ-வுக்கும் இருப்பதுபோல இ-க்கும் ஈ-க்கும் உறவிருக்கும்

ஆ-க்கும் அ-க்கும் இருப்பதுபோல ஈ-க்கும் இ-க்கும் உறவிருக்கும்

அ-க்கும் இ-க்கும் இருப்பதுபோல ஆ-வுக்கும் ஈ-க்கும் உறவிருக்கும்

இ-க்கும் ஆ-வுக்கும் இருப்பதுபோல ஈ-க்கும் ஆ-வுக்கும் உறவிருக்கும்

சிறப்பான விகிதங்கள் எண்களாகும்:

இருப்பினும் ஒத்த சொல் விகிதங்கள் பொதுவானதாகும். பின்வரும் எடுத்துக்காட்டுகள் இதனை வெளிப்படுத்தும்.

பாலூட்டிகளும் அவற்றின் இனம் உயிர்களும்:

பசு	:	கன்று
குதிரை	:	குட்டி

முட்டியிடுவனவும் அவற்றின் இனம் உயிர்களும்:

பாம்பு	:	குஞ்சு
பல்லி	:	குஞ்சு
மீன்	:	குஞ்சு
பறவை	:	குஞ்சு
முதலை	:	குஞ்சு

விலங்குகளும் ஒலியும்:

யானை	:	பிளிறுதல்
நரி	:	ஊளையிடுதல்
சிங்கம்	:	கர்சித்தல்
குதிரை	:	கனைத்தல்
கழுதை	:	கத்துதல்

செயலும் செய்பவரும்:

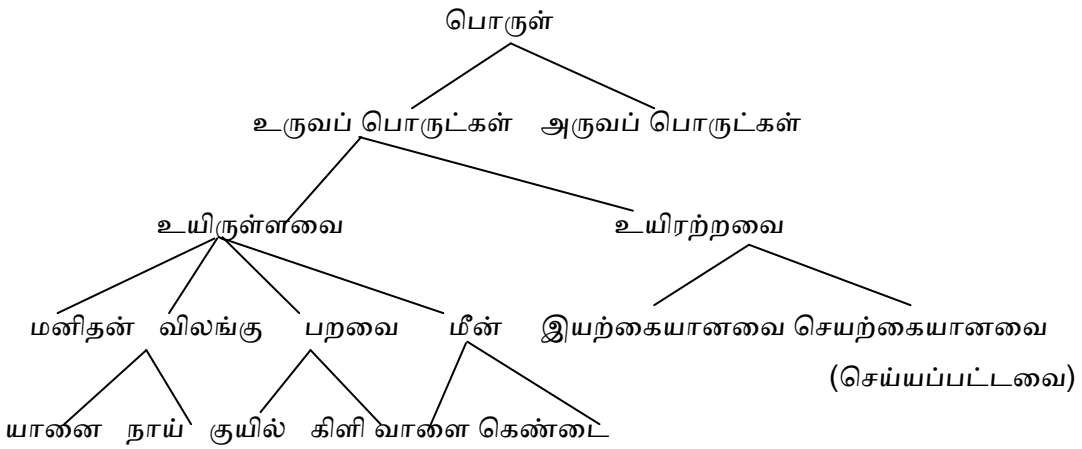
ஓட்டுதல்	:	ஓட்டுநர்
பெறுதல்	:	பெறுநர்
அனுப்புதல்	:	அனுப்புநர்

செயலும் கருவியும்:

சீவுதல்	:	சீப்பு
அரித்தல்	:	அரிப்பு

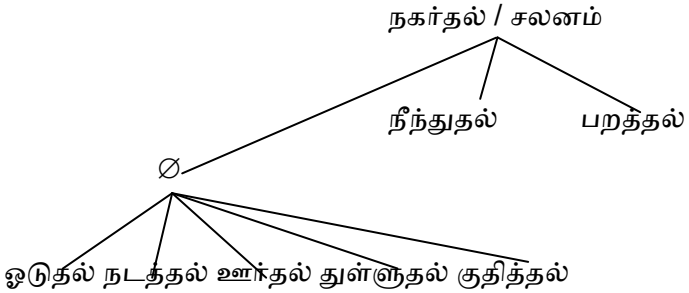
5.1.5.3. சொற்றொகுதியின் படிநிலை அமைப்பு

உள்ளடங்கு மொழியம் சொற்றொகுதியைப் படிநிலை அமைப்பாகக் (hierarchial structure of vocabulary) காண உதவுகிறது. சொற்றொகுதியோடல்லாமல் சொற்களங்களையும் (lexical fields) அது படிநிலை அமைப்பாக உணரச் செய்கிறது. இந்தப் படிநிலை உறவு அமைப்பைக் கிளைப்படங்கள் மூலம் காட்டலாம்.



யானைக்கும் விலங்கிற்கும் உள்ள உறவு உள்ளடங்கு உறவாகும். நாய்க்கும் பூனைக்கும் உள்ள உறவு இணக்கமற்ற உறவாகும் (incomplete relation).

இணை உள்ளடங்குமொழிகள் இணக்க உறவாக ஒன்றையொன்று மேலுறவாகவும் வரும்.

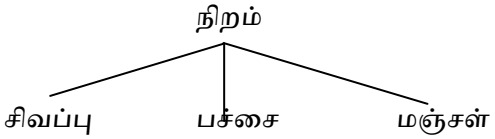


நிலத்தில் நிகழும் சலனத்தைக் குறிப்பதற்கு உள்ளடக்கும் சொல்லில்லை. இணை உள்ளடங்குச் சொற்கள் பெரும்பாலும் இணக்கம் அற்றவைகள். சில உள்ளடங்குச் சொற்கள் இணக்கமாகும். பல்லி, எலி, எறும்பு ஆகியன நடப்பதும் ஊர்வதும் ஒரே மாதிரியான செயல்தான். தவலை துள்ளுவதும் கங்காரு குதிப்பதும் ஒரே மாதிரியான செயல்தான். சில சமயங்களில் ஓடுவதும் பறப்பதும் ஒன்றாகும்.

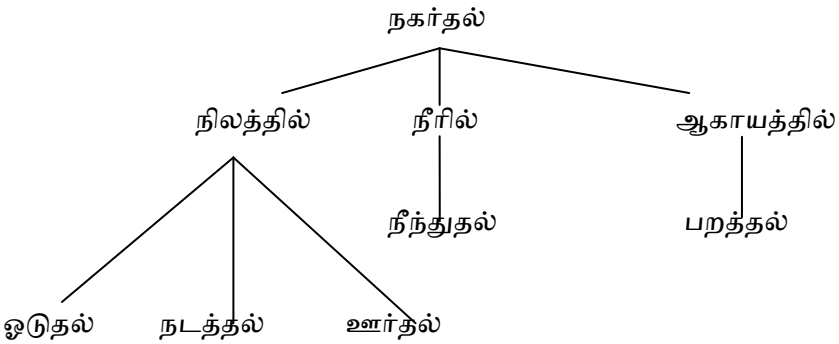
ஒரு மொழியின் மொத்தச் சொற்றொகுதியையும் உள்ளடங்கு, இணக்கமின்மை ஆகிய உறவுகளின் அடிப்படையில் படிநிலை அமைப்பாக அமைக்க முடியுமா என்ற வினா எழுவது நியாயம். அது மட்டுமல்லாமல் ஏதாவது சொல்வகையையோ (part of Speech) அல்லது அதைவிட குறைந்த நிலையில் சொற்களங்களையோ படிநிலையில் காண இயலுமா என்ற வினா எழலாம். சொல்வலையில் இத்தகைய சிக்கல்கள் தீர்க்கப்பட்டுள்ளன.

பெயர்ச்சொற்கள் யாவற்றையும் உள்ளடக்கிய உள்ளடக்குமொழிகளோ வினைச் சொற்கள் யாவற்றையும் உள்ளடக்கிய உள்ளடக்குமொழிகளோ மொழியில் இல்லை. பெரும்பாலான சொல் இரட்டைகளையோ சொற்கூட்டங்களையோ சொற்களங்களையோ உள்ளடக்கிய உள்ளடக்கு மொழிகள் கூட இல்லை. ஒரு உள்ளடக்குமொழிகளின் கீழ்வரும் உள்ளடங்குமொழிகள் இணக்கமற்று வேறுபட்டு இருப்பதில்லை. அவை பெரும்பாலும் இணக்கமாக மேலுறலாக இருக்கும். ஒரு சொல்லுக்குப் பல பொருள்கள் இருப்பதால் சொற்களைப் படிநிலை அமைப்பில் தருவதில் சிக்கல் ஏற்படுகிறது. பின்வரும் வழிகளை நாடலாம்.

அ. உள்ளடக்குமொழிகள் இல்லாத போது உள்ளடங்குமொழிகளை உள்ளடக்கும் அபூர்வ உள்ளடக்குமொழிகளைப் பயன்படுத்தலாம்.

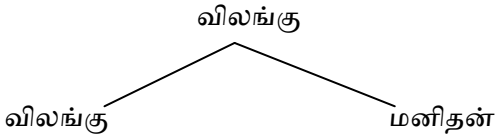


ஆ. சொற்களை படிநிலை அமைப்பாக பிரித்து வரும்போது சொல்லுக்குப் பதிலாகப் பொருண்மை பண்புக் கூறைப் பயன்படுத்தலாம்.

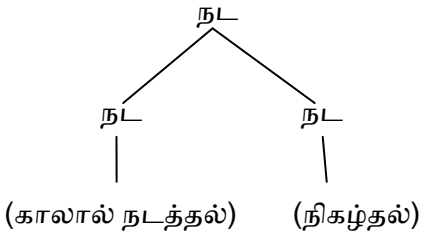


இப்படி முழுமையற்ற உள்ளடங்குமொழிய உறவை முழுமையான உள்ளடங்குமொழிய உறவுடன் கலந்து கையாண்ட ஒரு மொழியில் சொற்றொகுதியைப் படிநிலை அமைப்பாக மாற்றிக் காண்பது கைக்கூடும்.

இ. உள்ளடங்குமொழிகள் இணக்கமாக மேலுறவாக இருப்பின் அதன் அடிப்படை பொருளைப் பிரித்துச் சொல்லை தேவைக்குத் தகுந்தபடி இரண்டாகவோ மூன்றாகவோ காட்டலாம்.

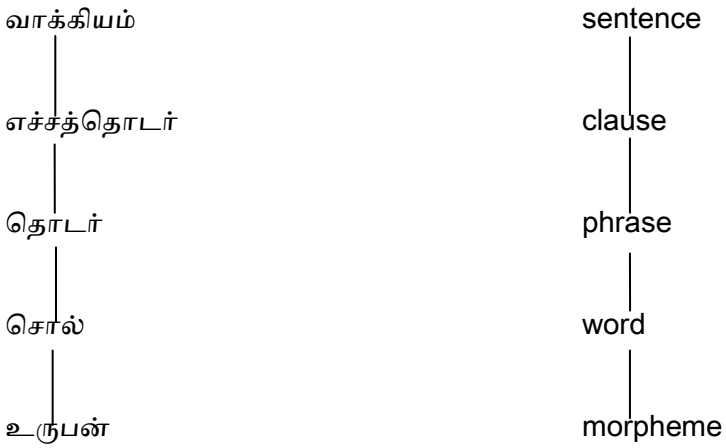


ஒருசொல் பலபொருள் சொற்களையும் இவ்வாறு பிரிக்கலாம்.

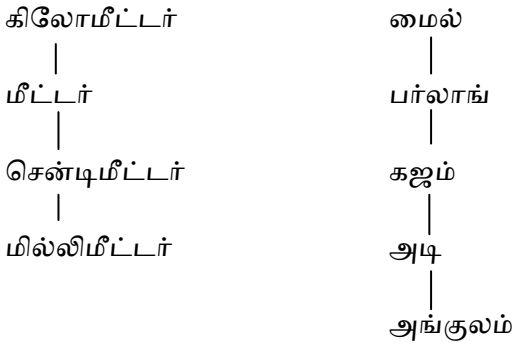


கிளைகளற்ற படிநிலை அமைப்பு (Non-branching Hierarchical structure)

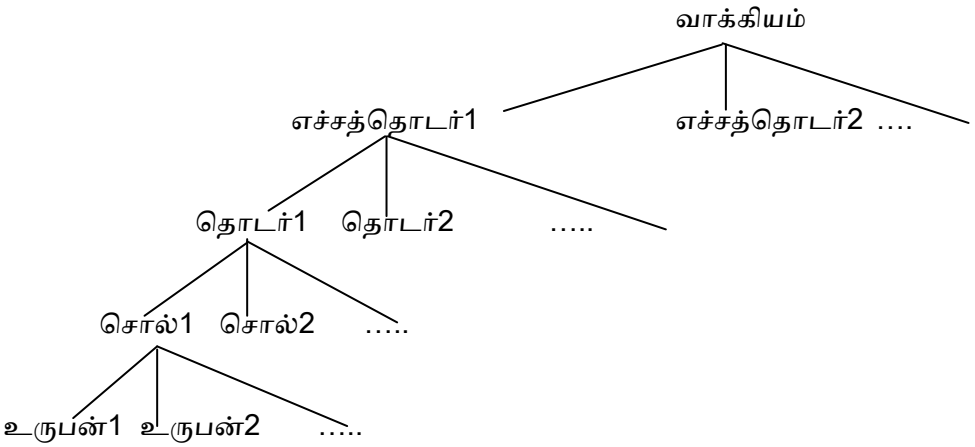
கிளைகளகற்ற படிநிலையாக வரும் சொற்கள் பல சொற்றொகுதியில் காணப்பெறும்.



மேற்சொன்ன எடுத்துக்காட்டிலுள்ள சொற்கள் வாக்கியத்திலிருந்து உருபன் வரையுள்ள பல நிலைகளைக் காட்டும் இலக்கணச் சொற்களாகும். இச்சொற்றொகுதியைக் கிளைகளால் காட்ட முடியும் என்றாலும் அதற்கு அவசியம் இல்லாமல் போய்விடுகிறது. ஒரு வாக்கியத்தைப் பல எச்சத் தொடர்களாகவும் எச்சத் தொடர்களைப் பல தொடர்களாகவும் தொடர்களைப் பல சொற்களாகவும் சொற்களைப் பல உருபன்களாகவும் விரிக்கலாம் என்றாலும் அப்படி பிரித்துக் காட்டுவது தேவையில்லை.



மேற்சொன்ன எடுத்துக்காட்டுகளில் ஒரு சொற்றொகுதியில் அடங்கும். சொற்கள் கிளைகளாக வர வாய்ப்புண்டு.



பல சொற்றொகுதிகள் எந்தவித கிளைகளற்றும் வருவன.

சொற்சங்கிலிகள் (chains), சொற்சுருள்கள் (helices) சொற்சக்கரங்கள் (cycles)

கிளைகளற்ற படிநிலை அமைப்பில் வரும் சொற்றொகுதியில் அடங்கும் சொற்களுக்கு இடையில் ஒரு ஒழுங்கு காணப்படும். இவ்வொழுங்கு அவைகளிடத்தில் இயற்கையாகவே காணப்படும். இந்த ஒழுங்கு முறைப்படி சொற்களை தொடர்ச்சியாக அமைக்க இயலும்.

நதி: ஆறு: வாய்க்கால் (x யானை: புலி: பூனை)

மேற்சொற்றொகுதியில் அடங்கும் சொற்கள் ஒரு ஒழுங்கில் வருவது மட்டுமல்லாமல் ஒரு படிதரத்தைக் காட்டுவதாக அமைகின்றன. சில சொற்றொகுதிகளில் வரும் சொற்களுக்கிடையில் படித்தரம் காணப்படுவதில்லை.

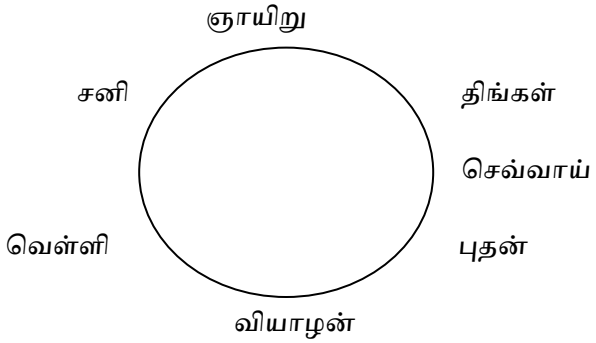
ஞாயிற்றுக்கிழமை: திங்கட்கிழமை: …

சொற்றொகுதியில் அடங்கும் சொற்கள் நீட்சியாக வரலாம், அல்லது சுற்றாக வரலாம். தோள்:மேற்கை:முழங்கை:முன்னங்கை:மணிக்கட்டு: உள்ளங்கை (சொற்சங்கிலி).

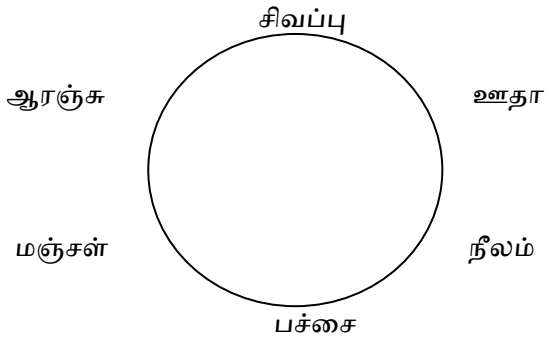
வாரத்தின் நாட்களைச் சங்கிலித் தொடராகவோ சக்கரத் தொடராகவோ காணலாம். ஞாயிற்றுக்கிழமையை வாரத்தின் முதல் நாளாகவும் சனிக்கிழமையை வாரத்தின் கடைசி நாளாக கருதுவோமானால் அவற்றைச் சங்கிலித் தொடர் அமைப்பாகக் காணலாம்.

ஞாயிற்றுக்கிழமை: திங்கட்கிழமை: செவ்வாய்கிழமை: …

அவ்வாறல்லாமல் ஒன்றையொன்று முடிவில்லாமல் தொடர்ந்து வருவதாகக் கருதுவோம். அவை சக்கரத் தொடராக அமையும். இவ்வாறு நீட்சியாகவும் சுற்றாகவும் வருவது சுருள் எனப்படும்.



இதுபோல் நிறங்களையும் சக்கரத் தொடராகக் காணலாம். ஆனால், நிறங்களுக்கும் கிழமைகளுக்கும் உள்ள வேறுபாடு என்னவென்றால் நிறங்கள் உண்மையான சக்கரத் தொடருக்கு எடுத்துக்காட்டாகும்.



நிறங்கள் படிநிலையாக வராது: இவற்றின் அமைப்பு உறவுகளுக்கு திசைநோக்கு குணம் கிடையாது: மேலும் கீழும் இல்லை. ஆனால் கிழமைப் பெயர்கள் உண்மை சக்கரத் தொடராக அமையாது. *ஞாயிறு, திங்கள், செவ்வாய், புதன், வியாழன்* என்ற அமைப்பில் முதலும் இறுதியுமானச் சொற்கள் ஒரே நாளைக் குறிக்காது.

நீட்சித் தன்மையும் சுற்றுத் தன்மையும் கலந்த பண்பு சொற்களுக்கு உண்டு. சொற்சுருளைச் சார்ந்த சொற்களுக்கு தெளிவான எல்லை உண்டு அல்லது தெளிவற்ற நிலையுண்டு.

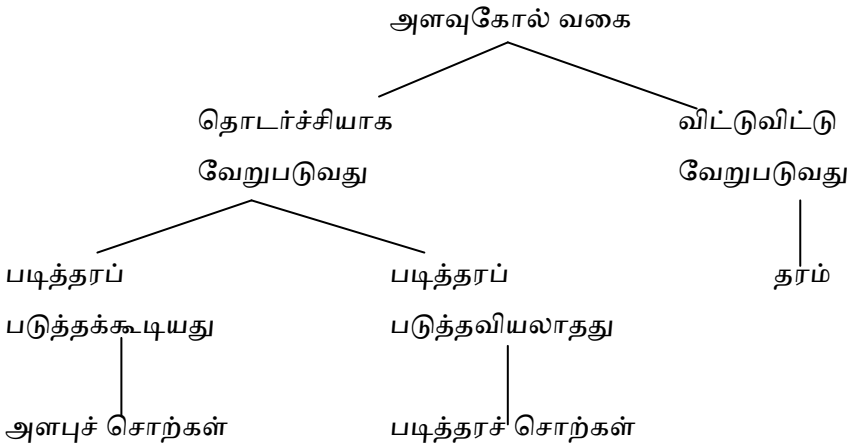
திங்கட்கிழமை : செவ்வாய்கிழமை மதியம் : சாயங்காலம்

ஜூன் : ஜூலை இலையுதிர்காலம் : குளிர்காலம்

சொற்சுருளில் காணப்படும் தெரிவுகள் காரணமற்றதல்ல: பலவற்றில் இயற்கையும் இடுகுறியும் காணப்படும்.

தரம் (ranks), படித்தரம் (grades), அளபு (degree)

இரண்டு வேறுபட்ட அளவுகோல் வகைகள் உண்டு. அவற்றில் ஒன்று தொடர்ச்சியாக வேறுபடுவது, மற்றொன்று சிறிது விட்டு விட்டு வேறுபடுவது. அளவுகோலில் செயலாற்றும் சொற்கள் தரம் எனப்படும். தொடர்ச்சியாக அளவுகோலில் வேறுபடுபவற்றைப் படித்தரப்படுத்தக்கூடியது, படித்தரப்படுத்த முடியாது என பிரிக்கலாம்.



தரங்கள் (ranks)

பல மட்டங்களைக் குறிக்கும் இராணுவ படிநிலைகள் தரங்களுக்குச் சிறந்த எடுத்துக்காட்டுகளாகும்.

கர்னல்: லெப்டினெட்: கர்னல்: மேஜர்: கேப்டன்: ……..

இதில் வேறுபடக்கூடிய அகநிலைப் பண்பு தரம் ஆகும். இப்பண்பு தொடர்ச்சியாக வேறுபடுவதில்லை. ஒரு மேஜர் மற்ற மேஜரை விட உயர்ந்த தரத்தில் உள்ளவர் என்று கூறமுடியாது: அதாவது *அவன் கொஞ்சம் குறைந்த மேஜர்* என்று கூற இயலாது. எண்ணுப் பெயர்கள் இராணுவத் தரங்கள் போல விட்டுவிட்டு வேறுபடுகின்றன.

ஒன்று: இரண்டு: மூன்று: நான்கு: ஐந்து: ……

‘*கொஞ்சம் குறைந்த ஐந்து*’ என்று கூறமுடியாது.

சுருள் சங்கிலித் தொடரை ஒத்த தொடர்ந்த மீண்டும் உபயோகிக்கப்பட்டு பெரிய எண்ணுப் பெயர்களாகத் தோன்றுகின்ற இரண்டு வகையான எண்ணுப் பெயர்களாவன:

‘அலகுகள்’ ஒன்று, இரண்டு, ……. ஒன்பது

‘பத்துகள்’ இருபது, முப்பது, …….. தொன்னூறு

இது தவிர அளவைச் சொற்களை ஒத்த ஒருவகைச் சொற்களும் உண்டு. அவையாவன, நூறு, ஆயிரம், இலட்சம், கோடி போன்றனவாகும்.

அவன் ஆயிரக்கணக்கில் சம்பளம் வாங்குகிறான்.

அவள் கோடிக்கணக்கில் பணம் வைத்திருக்கிறாள்.

சில சொல்வகைகள் எண்களை உள்ளடக்கியவை. அவைகளையும் தரங்கள் எனலாம்.

முதலாவது, இரண்டாவது, …..

அளபுச் சொற்கள் (degree words)

fail, pass, credit, distinction என்பன அளபுச் சொற்களுக்கு எடுத்துக்காட்டுகளாகும். ஒரு மேஜர் மற்றொரு மேஜரை விட தரத்தில் உயர்ந்திருக்க முடியாது. ஆனால் pass, credit-ஐ விட நல்ல உண்மையாக இருக்கலாம். சில அளபுச் சொற்கள் காலத்தொடர்ச்சியை காட்டுபனவாக அமைகின்றன.

baby, child, adolescent, adult.

வாலை, தருணி, பிரவிடை, விருத்தை, பேதை, பெரும்பை, மங்கை, மடந்தை, அரிவை, தெரிவை, பேரிளம்பெண்

தூரத்தையும் நேரத்தையும் குறிக்கின்ற அளவை அலகுகள் பகுதி-முழுமை (part-whole) படிநிலையை உணர்த்தினாலும் அவற்றை ஒரு பரிமாணத்தால் தொடர்ச்சியான அளவுகோலில் தொடர்படுத்தி அளபுச் சொற்களாக வகைப்படுத்தலாம். பிற அளவுச் சொற்கள் அடிப்படை பண்பு வழி தொடர்ச்சியாக நீட்சி முன்னேற்றத்தைக் காட்டும்போது அளவை அலகுகள் ஏறுமுகமாக வரும். எடுத்துக்காட்டாக *ஒன்று, இரண்டு,* ... போன்றவையும் *ஞாயிறு, திங்கள்*, ... போன்றவையும் அவைகளுக்கு இடையில் அளவுகோலில் சமமான இடைவெளி காட்டுகின்றன. *வினாடி:நிமிடம்:மணி: நாள், மில்லிமீட்டர், சென்டிமீட்டர், மீட்டர், கிலோமீட்டர்* என்பன ஏறுமுகமாக வருவன அல்ல.

படித்தரச் சொற்கள்

படித்தரச் சொற்களை நிரல்படுத்த முடிவதால் அவை அளபுச் சொற்களைவிட வேறுபடுகின்றன அவைகள் பெரும்பாலும் பெயரடைகளாகும்.

freezing (cold), cool, warm (hot), scorching.

கடுங்குளிர், குளிர், மிதவெப்பம், வெப்பம்

கீழ்வருவன படித்தரப்படுத்தக்கூடியது என்றாலும் சிறிது வேறுபட்டவை.

mound, hillock, hill, mountain

பாறை, குன்று, மலை

அவை உருவத்தில் வேறுபடுகின்றன. இவைகளைப் படித்தரப்படுத்தக்கூடிய சின்ன, பெரிய, மிகப்பெரிய/*small, large, huge* என்ற சொற்களுடன் தொடர்புபடுத்திக் கூறுவது கடினம்.

moderately large, fairly large, very large, extremely large என்ற சொற்களுடன் தொடர்புபடுத்தலாம்.

5.2. தமிழில் சொற்பொருண்மை மயக்கநீக்க முயற்சிகள்

இப்பகுதியில் தமிழ்த் தரவுத்தொகுதிகளில் சொற்பொருள் மயக்க நீக்கத்திற்கான முயற்சிகள் பற்றி விளக்கப்பட்டுள்ளது. குறிப்பாகப் பாஸ்கரன் அவர்களும் அருண் அவர்களும் செய்த ஆய்வு முயற்சிகள் சுருக்கமாகக் கூறப்பட்டுளது.

5.2.1அருணின் சொற்பொருண்மை மயக்கநீக்க முயற்சி

அருண் “Evaluation of Word Sense Disambiguation methods applied on Tamil Corpus” என்ற தமது முனைவர் பட்ட ஆய்வில் பல சொற்பொருண்மை மயக்கநீக்க நெறிமுறைகளைப் பட்டியலிட்டு அவற்றை மதிப்பீடு செய்கின்றார். அவருடைய ஆய்வின் நோக்கம் தமிழ் தரவுத்தொகுதியில் சொற்பொருண்மை மயக்கநீக்கத்தின் வேறுபட்ட நெறிமுறைளை மதிப்பீடு செய்வதாகும்.

5.2.1.1 நெறிமுறை

சொற்களுக்குப் பல அர்த்தங்கள் அல்லது பொருண்மைகள் உள்ளன. ஒரு சொல்லுக்குச் சரியான பொருண்மை தரப்பட்டுள்ள சூழலில் தரப்படவேண்டும். ஒரு சொல்லுக்கு எல்லைக்குட்பட்ட எண்ணிக்கையிலான வேறுபட்ட அர்த்தங்கள் இருப்பதாகக் கருதலாம். வழக்கமாக ஒரு குறிப்பிட்ட சொல் பயன்படுத்தப்படும் சூழல் எல்லைக்குட்பட்ட எண்ணிக்கையிலான பொருண்மைகளுக்குள் ஒரு பொருண்மையையோ அர்த்தத்தையோ தெரிந்தெடுக்கத் தடயத்தைத் தருகின்றது. அர்த்தங்கள் அகராதியிலும் சொற்களஞ்சியங்களிலும் பிற நோக்கீட்டு மூலவளங்களிலும் தரப்பட்டுள்ளன. சுருக்கமாகக் கூறினால் சொல்லின் பொருத்தமான அர்த்தம் அது பயன்படுத்தப்படும் சூழலில் இருந்தும் அகராதிகளிலிருந்தும் பெறப்படும்.

சொற்பொருண்மை மயக்கநீக்கம் தற்போது மிக முக்கியமான தலைப்பாகும். இது சொல்லின் சரியான அர்த்தத்தைக் கண்டுபிடிக்கவும் பொருண்மை மயக்கம் உள்ள சொல் வரும் வாக்கியங்களைப் பொருண்மை மயக்கநீக்கம் செய்யவும் ஒரு உபாயத்தை உருவாக்குவதையும் நோக்கமாகக் கொண்டுள்ளது. இப்பிரச்சனை இயற்கை மொழி ஆய்வுக்கு முக்கியமாகும்; குறிப்பாக வாக்கியப் பப்பாய்வுக்கு முக்கியமானதாகும். வழக்கமாகச் சொல்லின் பொருண்மை

மயக்கத் தன்மை காரணமாக ஒரு வாக்கியத்திற்குப் பல பகுப்பாய்வுகள் கிடைக்கப்பெறும்; இருப்பினும் ஒரே ஒரு வெளியீடு மட்டுமே சரியானதாக இருக்க இயலும். இச்சூழலில் சரியான வெளியீட்டைப் பெற சொல்லின் சரியான அர்த்தத்தைக் கண்டுபிடிக்கும் ஒரு கருவிதேவை.

5.2.1.2 சொல்மயக்கநீக்க நெறிமுறைகள்

அடிப்படையில் சொற்பொருண்மை மயக்கநீக்கத்திற்கு இரண்டு நெறிமுறைகள் உள்ளன: கண்காணிக்கப்பட்ட அணுகுமுறை 2. கண்காணிக்கப்படாத அணுகுமுறை. கண்காணிக்கப்பட்ட நெறிமுறை அர்த்தத்திற்கு அடையாளப்படுத்தப்பட்ட பனுவல்களை வேண்டும்; இதில் மனித அகராதியிலாரால் பல்பொருளொருமொழிகள் அதன் பொருண்மையால் புலக்குறிப்பு செய்யப்பட்டிருக்கும். இயந்திரம் கற்கும் உபாயங்கள் பகுப்பானை ஊக்குவிக்கப் பயன்படுத்தப்படும். இந்த அணுகுமுறை சிக்கனமானது அல்ல; ஏனென்றால் அடையாளப்படுத்தப்பட்ட பனுவல்கள் கிடைப்பது சிரமம். கண்காணிக்கப்படாத நெறிமுறை எந்த கற்றலையும் செய்வதில்லை மற்றும் அது அர்த்தம் அடையாளப்படுத்தப்பட்ட பனுவலை வேண்டாது. எனவே அது சிக்கனமானது.

5.2.1.2.1 சொற்பொருண்மை மயக்கநீக்கத்திற்கு பெய்சின் வகைப்படுத்தி

இந்நெறிமுறை கேல், சர்ச், யரோவ்ஸ்கி (1992) என்போரால் முன்மொழியப்பட்டது. இந்த நெறிமுறையில் இலக்குச் சொல்லின் வேறுபட்ட அர்த்தங்கள் அதன் சூழலுடன் பொருத்தப்படும்; சூழல் அடிப்படையில் சரியான அர்த்தம் அதற்குத் தரப்படும்.

5.2.1.2.2 அகராதி அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் (கண்காணிப்பற்றது)

இந்நெறிமுறை லெஸ்க் (Lesk 1986) என்பவரால் முன்மொழியப்பட்டது. இந்த நெறிமுறை பொருண்மை மயக்கநீக்கம் செய்ய விருப்ப வித்தாகச் சொற்களின் அகராதி வரையறை விளக்கத்தைப் பயன்படுத்துகின்றது.

5.2.1.2.3 தகவல் கோட்பாடுசார் அணுகுமுறை

இந்த நெறிமுறை பிரவுண், டெல்லா பிட்ரா, மெர்செர் (Brown, Della ietra, Mercer 1991) ஆகியோர்களால் முன்மொழியப்பட்டது. 'ஃபிளிப்-ஃப்ளாப் வழிமுறை வரைவு' என்று அழைக்கப்படும் இந்த நெறிமுறை இலக்குச் சொல்லின் எந்த அர்த்தம் பயன்படுத்தப்பட்டுள்ளது என்பதைக் கூறும் ஒரு சூழல் சொல்லையோ பண்புக்கூறையோ சார்ந்திருக்கும்.

5.2.1.2.4 கண்காணிக்கப்படாத வழிமுறை வரைவு கற்றல் (Unsupervised learning algorithm)

இந்நெறிமுறை யரோவ்ஸ்கியால் (Yarowsky 1995) முன்மொழியப்பட்டது. இந்நெறிமுறை விரிந்த 'சொற்களின் பைகள்' சூழலைப் பயன்படுத்துகின்றது; இருப்பினும் அண்மையில் வரும் சொற்களிலிருந்து இலக்குச் சொல்லின் அர்த்தத்திற்குத் தடயத்தைப் பெற 'ஒரு சேர்ந்துவருகைக்கு அல்லது சூழலுக்கு ஒரு அர்த்தம்' மற்றும் 'ஒரு கருத்தாடலுக்கு ஒரு அர்த்தம்' என்ற ஒருமொழிப் பண்புகளை உட்படுத்துகின்றது. அர்த்தம் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிக்குத் தேவையில்லை.

5.2.1.2.5 வாஸ்ப்-பெஞ்

இந்த நெறிமுறை கில்காரிஃப் மற்றும் டக்வெல் (Kilgarriff andTugwell 2001) முன்மொழியப்பட்டது. இந்நெறிமுறை மொழியியல் மற்றும் புள்ளியியல் தகவல்களை இணைக்கின்றது. ஒழுங்குமுறை சொல் வரும் சூழலைப் பார்க்கும், பொருத்தமான தொடர்புள்ள விதியைத் தெரிந்தெடுக்கும், மற்றும் பொருத்தமான அர்த்தத்தைத் தரும்.

5.2.1.2.6 மதிப்பீடு

முன்னர் கூறப்பட்ட நெறிமுறைகள் யாவும் தமிழ்த் தரவுத்தொகுதியில் பயன்படுத்தப்படும். இந்நெறிமுறைகளின் சரியான தன்மை சில மதிப்பீடு செயல்முறைகளைப் பயன்படுத்திச் செய்யப்பட்டுள்ளது.

5.2.2 பாஸ்கரனின் சொற்பொருண்மை மயக்கநீக்க ஆய்வு

பாஸ்கரனின் ஆய்வு (Baskaran 2002) சொற்பொருண்மை மயக்கநீகத்திற்கு முன்னோடியான ஆய்வாகும். அவர் தமது எம்.எஸ். பட்டத்திற்காக இவ்வாய்வை மேற்கொண்டார். இவ்வாய்வேட்டின் நோக்கம் ஒரு பனுவலில் உள்ள சொல் மயக்கங்களை நீக்க ஒரு பொருண்மை ஆய்வி உருவாக்குவதாகும். இந்த ஒழுங்குமுறைத் தமிழை மூலமொழியாகக்கொண்ட இயந்திர மொழிபெயர்ப்புக்கும் தமிழ் தரவுகளின் அறிவாற்றல் உள்ள தகவல் மீட்பு ஒழுங்குமுறைக்கும் தேவையானதாகும். எடுத்துக்காட்டாகத் தமிழ்-ஆங்கில இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை பின்வரும் வாக்கியத்தை உள்ளீடாகக் கொள்கின்றது என்று கொள்வோம்.

நீலகிரி காடுகளில் அரிய வகை விலங்குகள் காணப்படுகின்றன.

மேற்கண்ட வாக்கியத்தில் *விலங்குகள்* என்பது *விலங்கு* என்பதன் பன்மை வடிவமாகும். *விலங்கு* என்பதற்கு இரண்டு வேறுபட்ட அர்த்தங்கள் உள்ளன:

1. மிருகம்
2. கையில் போடும் விலங்கு

‘விலங்கு’ அர்த்தம் ‘(கை)விலங்கு’ அர்த்தத்தைவிட அதிக நிகழ்வெண்ணில் வருகின்றது. ஆங்கிலத்தில் இந்த இரு அர்த்தங்களுக்கும் தனித்தனியான வடிவச் சொற்கள் உள்ளன: விலங்கு1 ’animal’, விலங்கு2 ‘hand-cufs’. எனவே பொருண்மையியல் ஆய்வி இப்பொருண்மை மயக்கத்தை நீக்க வேண்டும்; இதன் படி மேற்சொன்ன தமிழ் வாக்கியம் பின்வருமாறு மொழிபெயர்க்கப்பட வேண்டும்.

Rare animal species are found in Nilgiris forests.

5.2.2.1 அணுகுமுறை

இவ்வாய்வில் பொருண்மை மயக்க நீக்கம் கொத்தாக்க உபாயம் (clustering technique) மற்றும் அறிவு மூலவளங்களாகச் சேர்ந்துவருகைகளையும் (collocations) வேற்றுமை குறியீடுகளையும் பயன்படுத்துவது அடிப்படையில் நடைமுறை படுத்தப்பட்டுள்ளது. ஒழுங்குமுறையின் செயல்பாடு படத்தில் தரப்பட்டுள்ள ஒழுங்குமுறை ஒழுக்குப் படத்தில் காட்டப்பட்டுள்ளது.

இவ்வணுகுமுறை இரண்டு கட்டங்களில் செயல்படுகின்றது. பயிற்சி கட்டத்தின் தொடக்கத்தில் ஒழுங்குமுறை பொருண்மை மயக்கம் உள்ள சொற்களைக் கண்டுபிடிக்கின்றது. வேறுபட்ட ஆய்வாளர்கள் வேறுபட்ட அளவுகளில் சூழல்களைப் (அதாவது வேறுபட்ட சாளர அளவுகளில் (different window size)) பயன்படுத்தியுள்ளனர். சாளர அளவு ஒழுங்குமுறையின் செயல்பாட்டைக் கணிசமான அளவில் பாதிக்கின்றது. ஒரு பெரிய சாளரம் சூழலுக்கு அதிக அளவில் தொடர்பில்லாத சொற்களைக் கொண்டுவருகின்றது; ஆனால் சிறிய சாளரம் சில முக்கியமான சேர்ந்துவருபவைகளை இழக்கின்றது. யரொவ்ஸ்கி (Yarowsky 1995) உட்பட பல சொற்பொருண்மை மயக்கநீக்கிகளில் காணப்படும் பொதுவான நடமுறையைப் பின்பற்றி இவ்வாய்வில் இருபது-சொல் சாளரம் பயன்படுத்தப்பட்டுள்ளது.

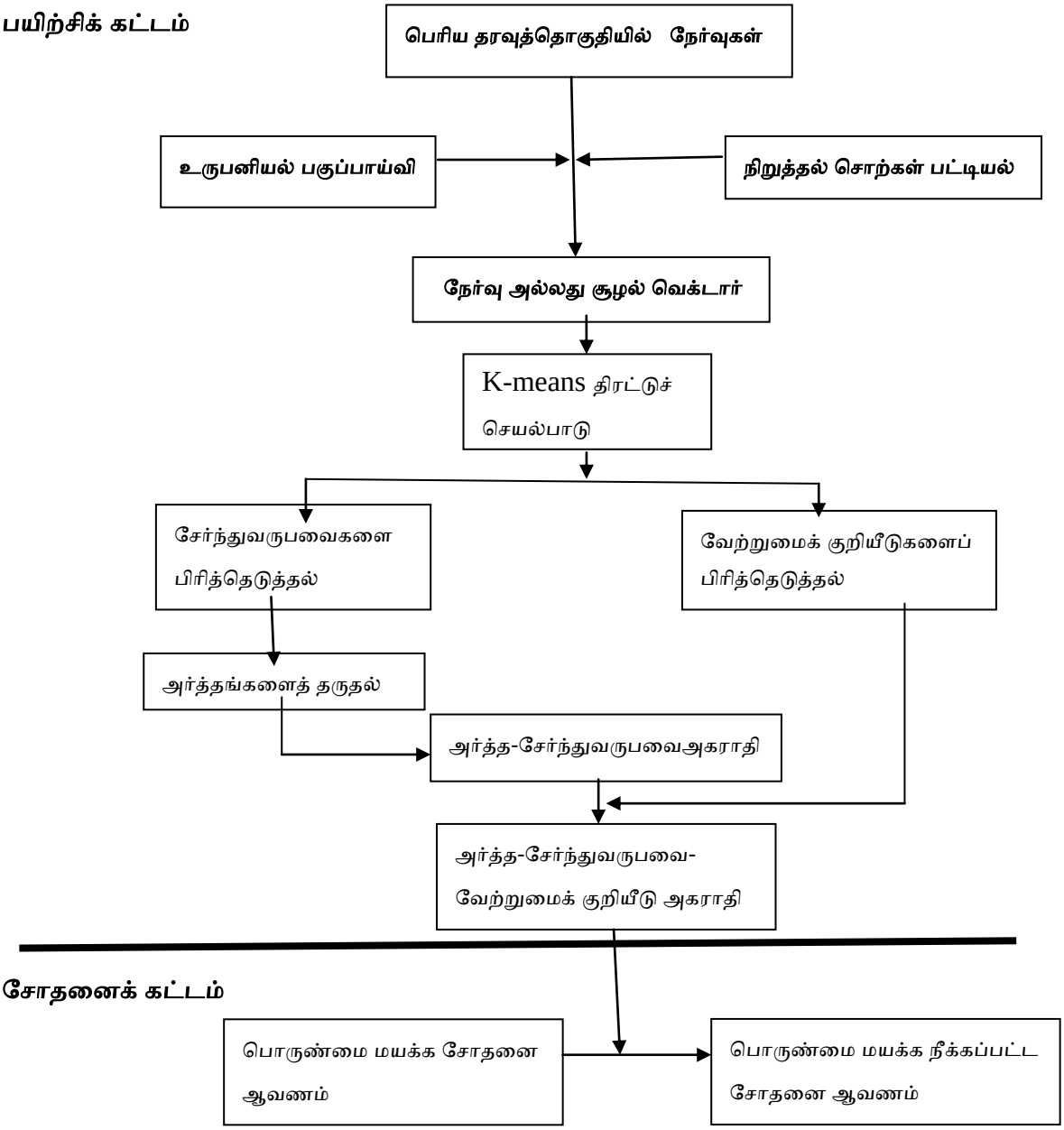
நிறுத்தல்-சொற்கள் (stop-words) என்று அழைக்கப்படுகின்ற அதிக நிகழ்வெண் உள்ள சொற்கள் சாளரத்திலிருந்து நீக்கப்படும். பின்னர் திரட்டுவதில் பயன்படுத்தப்படும் சூழல் சொற்களின் வேர்வடிவைப் பெற உருபனியல் அடிப்படையில் ஆயப்படும். ஒவ்வொரு சூழலும் சூழல் அல்லது நிகழ்வு வெக்டராகப் உருப்படுத்தம் செய்யப்படும். சேர்ந்துவருபவைகள்

தானியக்கமாக இந்தத் திரட்டுகளிலிருந்து சேகரிக்கப்படும் மற்றும் அர்த்தங்கள் மனித-அடையாளாப்படுத்திகளால் (human-annotators) ஒதுக்கப்படும்/தரப்படும்; இதன்படி ஒரு அர்த்த-அகராதி உருவாக்கப்படும். வேற்றுமை-குறியீடுகள் ஒவ்வொரு திரட்டுக்கும் அடையாளம் காணப்படும் மற்றும் அர்த்த-சேர்ந்துவருபவை வேற்றுமை-குறியீடு அகராதியை (sense-collocation case-marker dictionary) உருவாக்க வேண்டி சேர்க்கப்படும்.

இரண்டாவது கட்டத்தில் ஒழுங்குமுறை பொருண்மை மயக்கநீக்கத்திற்கு தயாராக இருக்கும். பொருண்மை மயக்கச் சொற்கள் கொண்ட எந்தப் பனுவலும் ஒழுங்குமுறைக்குத் தரப்படலாம். படத்தில் காட்டியுள்ளது போன்று ஒழுங்குமுறை பொருண்மை மயக்கச் சொற்களின் பார்க்கப்படாத நிகழ்வுகளை பொருண்மை மயக்கநீக்கம் செய்ய அர்த்த-சேர்ந்துவருகை வேற்றுமை-குறியீட்டு அகராதியைப் (Sense-Collocation-Case marker dictionary) பயன்படுத்துகின்றது. ஒவ்வொரு பொருண்மை மயக்கம் உள்ள சொற்களுக்கும் பயிற்சி ஒருநேரச் செயல்பாடாகும். ஒழுங்குமுறை ஒரு குறிப்பிட்ட சொல்லுக்காகப் பயிற்சி பெற்றுவிட்டால் அதை ஒரு தரப்பட்ட சூழலில் அச்சொல்லின் பொருண்மை மயக்கத்தை நீக்கப் பயன்படுத்த இயலும். இந்த நெறிமுறையில் சேர்துவருபவைகளும் வேற்றுமைக் குறியீடுகளும் ஒரு முக்கியமான பங்களிப்பு செய்கின்றது என்பதைக் காணலாம்.

ஒழுங்குமுறை ஒழுக்குப் படம்

பயிற்சிக் கட்டம்



சோதனைக் கட்டம்

5.2.2.1.1 சேர்ந்துவருபவைகளின் முக்கியத்துவம்

சேர்ந்துவருபவைகள் தந்துள்ள ஒரு வாக்கியத்தில் பொருண்மை மயக்கச் சொற்களின் அர்த்தத்தை உறுதியாக அனுமானிக்கும் அடுத்துவரும் சொற்கள் ஆகும். பொதுவாகச் சேர்ந்துவருபவை இரண்டு சொற்களுக்கு இடையில் அளக்க இயலும் இட-சிறப்பு உறவுகளைக் குறிப்பிடுகின்றது. சேர்ந்துவருபவைகள் பொருண்மை மயக்கமுள்ள சொற்களுக்குப் பொருண்மையியல் அடிப்படையில் மிக அண்மையில் வரும் சொற்களைப் பற்றிய குறியாக்கம் செய்யப்பட்ட தகவலாகும். எடுத்துக்காட்டாகப் பின்வருமாறு இரு அர்த்தங்கள் உள்ள plant என்ற சொல்லைக் கருத்தில் கொள்ளவும்.

plant1 – a living organism

plant2 – a manufacturing place

plant என்ற சொல்லின் சேர்ந்துவருபவைகள் பனுவலில் *plant* என்பதுடன் அடிக்கடிச் சேர்ந்துவரும் சொற்களை உட்படுத்தும். *plant* என்பதன் மேற்கண்ட அர்த்தங்களுக்கு எடுத்துக்காட்டான சேர்ந்துவருபவைகள் கீழே பட்டியலிடப்பட்டுள்ளன.

plant1 – growth, height, flower, fruit, species, leaves

plant2 – car, union, equipment, assembly, nuclear, job, worker

பெரும்பாலும் சேர்ந்துவருபவைகள் பொருண்மை மயக்கமுள்ள சொற்களிருந்து உள்ள தூரத்திற்கு உணர்வுள்ளவைகளாகும். பொருண்மை மயக்கச் சொற்களிலிருந்து குறைந்த தூரத்தில் உள்ள சொற்கள் வலுவாக அதன் அர்த்தத்தை சுட்டிக்காட்டும். தூரம் அதிகரிக்கும் போது பொருண்மைமயக்கச் சொற்களுடன் உள்ள அதன் சம்பந்தம் குறையும். தூரம் பொதுவாகச் சாளர அளவு என்று குறிப்பிடப்படும். இருப்பினும் சேர்ந்துவருவது என்று கருதுவதற்கு ஒரு சொல் பொருண்மைமயக்கம் உள்ள சொல்லிலிருந்து இருக்கவேண்டிய அறுதியான தூரம் இல்லை. சில வேளைகளில் சேந்துவருபவைகள் பொருண்மைமயக்கச் சொற்களிலிருந்து மிகுந்த தூரத்தில் வரலாம். தற்போதைய அணுகு முறையில் சேர்ந்துவரும் சொற்கள் யாவும் தானியக்கமாகத் திரட்டப்பட்ட நேர்வுகளிலிருந்து பிரித்தெடுக்கப்படும். இது விதை சேர்ந்துவருபவைகளைக் (seed collocations) கையால் தரும் தேவையை நீக்குகின்றது; ஒரு மொழியில் வரும் பொருண்மை மயக்கமுள்ள சொற்களைக் கருத்தில் கொள்ளும்போது இது சிரமமான வேலையாகும்.

5.2.2.1.2 வேற்றுமை குறியீடுகளின் முக்கியத்துவம்

ஆங்கிலத்தில் பெயருக்கும் வினைக்கும் உள்ள உறவுகள் *in, on. at, by, with* போன்ற முன்னுருபுகளால் வெளிப்படுத்தப்படும். தமிழில் இது பின்னுருபுகளால் வெளிப்படுத்தப்படும். *வர்ணம்* என்ற தமிழ்ச் சொல்லை எடுத்துக்கொள்வோம். இதற்கு குறைந்தது மூன்று வேறுபட்ட அர்த்தங்கள் உள்ளன.

1. நிறம்
2. இந்திய மரபு இசையில் ஒரு குறிப்பிட்ட பாடல் வகை
3. பண்டை இந்தியாவில் பிறப்பு அடிப்படையிலான பகுப்பு

இவ்வர்த்தங்கள் வாக்கியங்களில் வருகையில் வேற்றுமை உருபுகளின் தேர்வு சில வேளைகளில் இப்பொருண்மை வேறுபாடுகளை வெளிப்படுத்தலாம். *மாலை* என்ற சொல்லை எடுத்துக்கொள்வோம். இது மாலைக் காலத்தையும் கழுத்தில் அணியும் மாலையையும் குறிப்பிட்டுப் பொருண்மை மயக்கம் காட்டும். இவை பின்வருமாறு வாக்கியங்களில் பயன்படுத்தப்படலாம்.

மாலையில் பள்ளி ஆண்டுவிழா கொண்டாடப்பட்டது.

அவன் கழுத்தில் மாலை அணிவித்தார்கள்.

மேற்கண்ட வாக்கிய நேர்வுகளில் வேற்றுமை உருபுகள் அர்த்தங்களை வேறுபடுத்த உதவக்கூடும்.

5.2.2.2 கற்றல் கட்டம்

கொத்தாக்கல்/திரட்டல் செயற்பாங்கு (clustering process) இங்கு விளக்கப்பட்டுள்ளது.

5.2.2.2.1 சூழல்களும் சூழல் இடைவெளியும்

ஒரு ஆவணத்தில் சொற்பொருள் மயக்கமுள்ள சொல்லின் ஒவ்வொரு நேர்வும் அதன் சூழல் என்று அழைக்கப்படும். ஒரு பெரிய தரவுத்தொகுதியில் சொற்பொருண்மை மயக்கம் உள்ள சொல்லின் எல்லா சூழலகளும் S என்ற கருத்துச்சாயலால் குறிப்பிடப்பட்ட சூழல் இடைவெளியில் (context space) உருப்படுத்தம் செய்யப்படும். எடுத்துக்காட்டாக *மாலை* என்ற சொல்லைக் கருத்தில் கொள்வோம். 'மாலை (காலம்)' என்ற அர்த்தம் S1 என்றும் '(கழுத்திலிடும்) மாலை' என்ற அர்த்தம் S2 என்றும் கொள்வோம். மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட மூன்று மில்லியன் தமிழ்த் தரவுத்தொகுதியிலிருந்து பிரித்து எடுக்கப்பட்ட வாக்கியங்களில் வரும் *மாலை* என்பதை இரு அர்த்தங்களும் அடையாளப்படுத்தப்பட்டு கீழே தரப்பட்டுள்ளன.

தெரசாவின் அறப்பணிகளுக்கு நிகரில்லாத புகழ் மாலைகளாக அமைந்து மகிழ்வித்தன. S2

...தினந்தோறும் காலை மாலை ஆகிய இரு வேளைகளிலும் S1

காலையில் சூரியம் தோன்றிய நேரத்தில் இருந்து மாலையில் மறையும் நேரம் வரை S1

சூரியகாந்திப்பூ காலை, மாலை வேளைகளில் திசை S1

சுற்றுலாப் பேருந்து அடுத்த வெள்ளிக்கிழமை மாலை சென்னைக்குத் திரும்புகிறது S1

காலையில் 5 மணிக்கும், பிறகு 9:30 மணிக்கும், மலையில் 6:30 மணிக்கும் S1

அரசு தூக்கிலிட முடிவு செய்த நாளின் முதல் நாள் மலையே, தூக்கிலிட்டுக் S1

தத்தும் வீசிய துண்டறிக்கைகள் அன்று மாலையே "இந்துஸ்தான் டைம்ஸ்" S1

நடவடிக்கைகளின் காரணமாக நேற்று மாலையிலேயே வந்த செய்தியின்படி S1

"குரங்கு கையிலே பூமாலை" என்று அதைப்போல ஊராட்சி S2

சிலர், 'நான் மலையையே மலர் ஆக்குவேன் S2

அழகு, இளமை, கருவிழி, செல்லம், வாகனம், மாலை, கொடி முதலியவை S2

அல்லது திருநீறு காணப்படும் சபையோர் போடும் மாலைகளைப் கழற்றாமலே S2

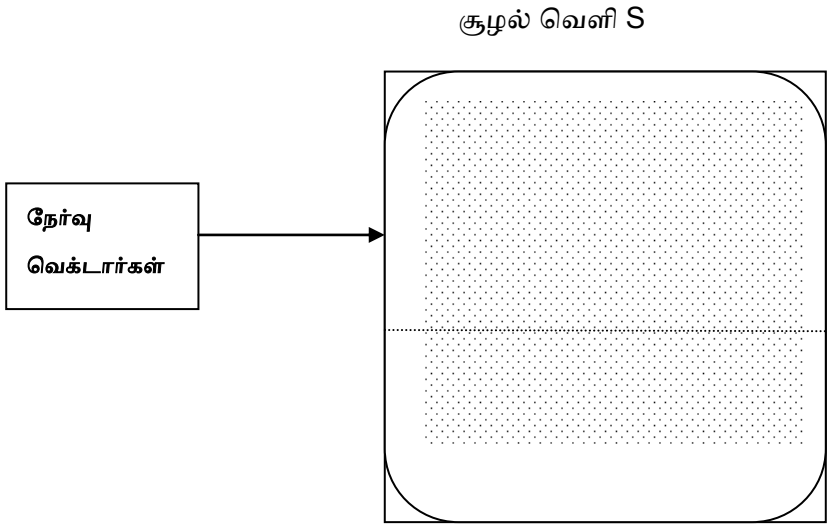
நானும் அந்தச் செய்தியை நேற்று மாலை அமைச்சரவைக் கூட்டம் S1

உணவுப் பொருள்களை உட்கொள்ளும் பொழுதும், மாலையில் மேய்ச்சலுக்குப் பிறகும் S1

'மாலை (காலம்)' S1என்ற அர்த்தத்தில் வந்த எல்லா சூழல்களும் படத்தின் மேல்பாகத்திலும் '(கழுத்தில் அணியும்) மாலை' S2 என்ற அர்த்தத்தில் வந்த எல்லா சூழல்களும் படத்தின் கீழ்ப் பகுதியிலும் காட்டப்பட்டுள்ளன. பிரிக்கக்கூடிய புள்ளிக் கோடு இவ்விரு அர்த்தங்களின் சூழல் வெளியைக் காட்டுவதற்காக இடப்பட்டுள்ளது. கொத்தாக்கச்/திரட்டுச் செயற்பாங்குகள் வேறுபட்ட சுழல்களின் உண்மையான பிரிவை எய்துகின்றது.

5.2.2.2.2. சூழல் இடைவெளியில் சூழல்களை உருப்படுத்தம் செய்தல்

சொற்பொருண்மை மயக்கம் உள்ள சொல்லின் ஒவ்வொரு சூழலும் சூழல் வெக்டார்களாகக் கையாள இயலும்; அவை சூழல் வெளியில் உருப்படுத்தம் செய்யப்படும். சூழல் வெக்டாரின் தனிமங்கள் பொருள்மயக்கம் உள்ள சொல்லின் எல்லா சூழல்களிலும் தோன்றும் எல்லா சொற்களையும் உட்படுத்தும். சூழல்களை சூழல் வெக்டார்களாக உருப்படுத்தம் செய்யும் நெறிமுறை ஒரு எடுத்துக்காட்டால் கீழே காட்டப்பட்டுள்ளது.



இப்போது பொருளடக்கச் சொற்களை அடையாளம் காண்பதன் செயற்பாங்கை விளக்க மூன்று சூழல்களைக் கருத்தில் கொள்ளவும்.

C1 – மாலையில் நடைபெற்ற விழாவில் மாணவர்களுக்குப் பரிசு வழங்கப்பட்டது

C2 – வரவேற்பு விழா மாலை சரியாக ஆறு மணிக்குத் துவங்கியது.

C3 – ரோஜாப்பூ மாலை இருபது ரூபாயில் கிடைக்கிறது

சூழல் சாளரத்தில் வரும் எல்லாச் சொற்களும் உருபனியல் அடிப்படையில் ஆயப்பட்டது மற்றும் விளையும் வேர்வடிவுகள் திறனுள்ள பொருளட்டக்கச் சொற்களாகக் கருதப்படும். இவை நிறுத்தல் சொற்களை நீக்கி மேலும் குறைக்கப்படும். C1-க்கு ஆறு பொருளடக்கச் சொற்களைக் கண்டுபிடிக்க இயலும்.

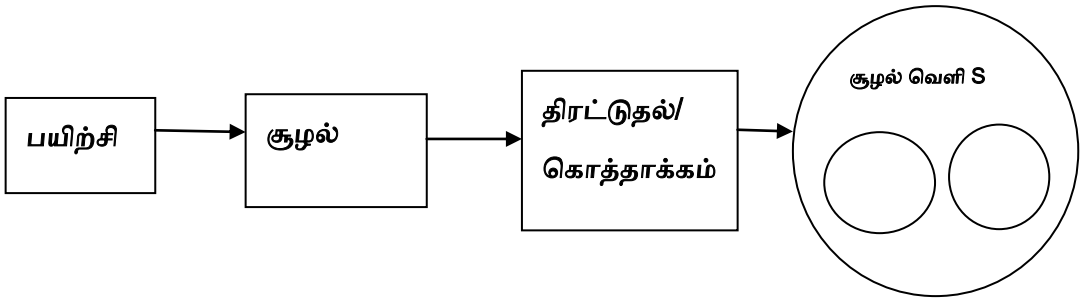
மாலை, நடைபெற்ற, விழா, மாணவர், பரிசு, வழங்கு

இதே வழியில் தரவுத்தொகுதியில் வரும் எல்லா நேர்வுகளுக்கும் பொருளடக்கச் சொற்களைக் கண்டுபிடிக்க இயலும். இவ்வாறு நமது குறைந்த மூன்று நேர்வுகளுக்கு 5 பொருளடக்கச் சொற்களைக் கண்டுபிடிக்க இயலும்.

மலை, நடைபெற்ற, விழா, மாணவர், பரிசு, வழங்கு, வரவேற்பு, ஆறு, மணி, துவங்கு, ரோஜாப்பூ, இருபது, ருபாய், கிடைக்கிறது

ஒவ்வொரும் நேர்வும் ஒரு சூழல் வெக்ட்டராக இப்போது உருப்படுத்தம் செய்யப்படும்; இதைச் சூழல் வெளியில் ஒரு தனிப்பட்ட இடத்தில் காட்ட இயலும். சூழல் வெக்டாரின் நீளம் தரவுத்தொகுதியில் எல்லா நேர்வுகளுக்கும் அடையாளம் கண்டுபிடிக்கக்கூடிய பொருளடக்கச் சொற்களின் எண்ணிக்கை ஆகும். இவ்வாறு எடுத்துக்காட்டாக *C1*, *C2* மற்றும் *C3* முன்று சூழல்களுக்கு ஒவ்வொரு சூழல் வெக்டாரின் நீளம் 14 ஆகும். சூழல் வெக்டாரில் உள்ள ஒவ்வொரு தனிமமும் அந்த குறிப்பிட்ட சொல் அச்சூழலில் இருக்கிறதா அல்லது இல்லையா என்பதைப் பொறுத்து பூஜியத்தையோ ஒன்றையோ எடுக்கும். திரட்டும் வழிமுறை வரைவு பொருண்மை மயக்கம் உள்ள சொல்லின் நேர்வுகளை வேறுபட்ட திரட்டுகளாகக் குழுமும்; ஒவ்வொரு கொத்தும்/திரட்டும் சூழல் வெளியில் ஒரு குறிப்பிட்ட இடத்தில் செறிவாக இருக்கும். கொத்தாக்கத்தின்/திரட்டுவதன் அடிப்படைக் கருத்து ஒரு கொத்தில்/திரட்டில் சூழல் அடிப்படையில் ஒற்றுமையுள்ள நேர்வுகளைக் குழுமுவதாகும். பின்வரும் படம் இவ்வணுகுமுறையில் கொத்தாக்கத்தின்/திரட்டும் செயற்பாங்கின் செயல்பாட்டைத் தெளிவாகக் காட்டுகின்றது. இது பொருண்மை மயக்கநீக்கதிற்கு வேண்டி இங்குப் பயன்படுத்தப்பட்டுள்ள திரட்டும் உபாயத்தை எடுத்துக்காட்டுகின்றது.

தரவுத்தொகுதியிலிருந்து பொருண்மை மயக்கச்சொற்களின் எல்லாச் சூழல்களும் முதலில் சூழல் வெக்டார்களாக உருப்படுத்தம் செய்யப்படும். பின்னர் சூழல் வெக்டார்கள் எல்லாம் K-மீன்ஸ் வழிமுறைவரைவைப் பயன்படுத்தித் திரட்டப்படும்.



கொத்தின்/திரட்டின் இறுதியில் சூழல் அடிப்படையில் ஒத்திருக்கின்ற கொத்துக்/திரட்டுக் குழுக்கள் பெறப்படும். கொத்துக்களிலிருந்து/திரட்டுகளிலிருந்து சேர்ந்துவருபவைகளைப்

பிரித்தெடுத்தபின் அடுத்த நடவடிக்கை அர்த்தம் அடையாளப்படுத்துவது ஆகும். இங்கு மனித அடையாளப்படுத்துபவர்கள் உயர்ந்த பத்து சேர்ந்துவருகைகளும் பொருண்மை மயக்கச் சொற்களின் அர்த்தங்களும் தரப்படுவார்கள். சொற்கள் வேறுபட்ட குழுமங்களாகப் பிரிக்கப்படும் மற்றும் ஒவ்வொரு குழுமமும் வேறுபட்ட அடையாளப்படுத்துபவரிடம் கொடுக்கப்படும். ஒவ்வொரு சேர்ந்துவரும் குழுமத்திற்கும் கூடுதல் பொருத்தமான அர்த்தத்தைத் தொடர்புபடுத்தி ஒரு அர்த்தம்-சேர்ந்துவருகை அகராதி உருவாக்குவதாகுவது அடையாளப்படுத்துபவர்களின் வேலையாகும்.

சொல்	**அர்த்தம்**	**சேர்த்துவரும் சொல்**
மாலை	மாலைப் பொழுது	காலை, பகல், சூரியன், மறை, நேரம், நாள், மாதம், மதியம், இரவு
மாலை	(கழுத்திலிடும்) மாலை	பூ, மணம், ரோஜா, மல்லிகை, மலர், பூஜை, மகளிர், கொன்றை
நூல்	நூல் (இழை)	இழை, தறி, சாயம், நெய், கதர், நெசவு, ஊடு, பாவு, துணி, ஊசி
நூல்	புத்தகம்	ஆசிரியர், பதிப்பகம், அச்சு, பக்கம், தாள்
விலங்கு	மிருகம்	காடு, சிங்கம், புலி, தோல், யானை, தந்தம், நாய், பூனை
விலங்கு	(கை) விலங்கு	குற்றவாளி, கோர்ட், போலீஸ், கைதி, விசாரணை, தண்டனை

வேற்றுமை குறியீடுகள் பெயர்களுக்கும் அல்லது பெயர்த்தொடர்களுக்கும் வினைகளுக்கும் அல்லது வினைத்தொடர்களுக்கு இடையே உள்ள உறவுகளை வெளிப்படுத்தும். பொருண்மை மயக்கம் உள்ள சொல்லின் அக்கம்பக்கத்தில் வரும் கொத்தாக்கப்பட்ட/திரட்டப்பட்ட நேர்வுகளிலிருந்து வேற்றுமைக் குறியீடுகளை அடையாளம் கண்டுப் பெயருக்கும் பிற பெயர்களுக்கும் அல்லது வினக்கும் உள்ள உறவுகளைக் கண்டறியலாம். ஒரு குறிப்பிட்ட கொத்தில்/திரட்டில் தோன்றும் எல்லா வேற்றுமைக் குறியீடுகளும் அடையாளம் காணப்படும் மற்றும் பின்னர் அர்த்தத்தம் சுட்டிக்காட்டும் முக்கியமான வேற்றுமைக் குறியீட்டைத் தேர்ந்தெடுக்க இப்பட்டியல் குறைக்கப்படும். முக்கியமான வேற்றுமைக் குறியீடுகளை

அடையாளம் காண்பதன் முழுச் செயல்பாடும் உருபனியல் பகுப்பாய்வியைப் பயன்படுத்தித் தானியக்கமாகச் செய்யப்படும். அடையாளம் காணப்பட்ட வேற்றுமை குறியீடுகள் அர்த்தம்-சேர்ந்துவருகை வேற்றுமைக் குறியீட்டு அகராதி உருவாக்க அர்த்தம்-சேர்ந்துவருகை அகராதியில் சேர்க்கப்படும். அகராதியில் மாதிரிப் பதிவுகள் வேற்றுமைக் குறியீடுகளைச் சேர்த்தபின் கீழ்வருமாறு அமையும்.

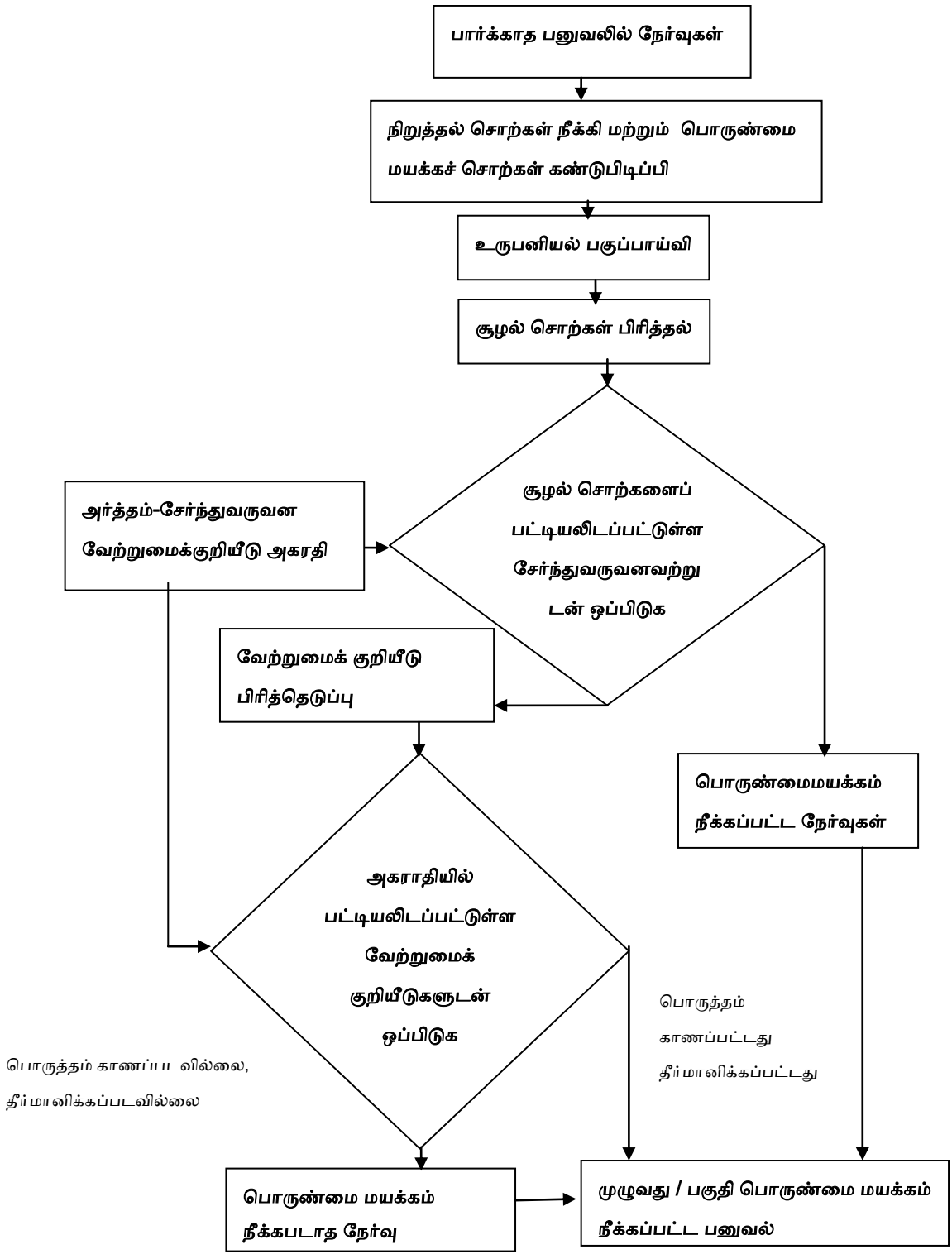
சொல்	**பொருள்/அர்த்தம்**	**சேர்ந்துவருபவைகள்**	**வேற்றுமை**
மாலை	மாலைப் பொழுது	காலை, பகல், சூரியன், மறை, நேரம், நாள், மாதம், மதியம், இரவு	இடவேற்றுமை
மாலை	(கழுத்திலிடும்) மாலை	பூ, மணம், ரோஜா, மல்லிகை, மலர், பூஜை, மகளிர், கொன்றை	கொடை வேற்றுமை
நூல்	நூல் (இழை)	இழை, தறி, சாயம், நெய், கதர், நெசவு, ஊடு, பாவு, துணி, ஊசி	இல்லை
நூல்	புத்தகம்	ஆசிரியர், பதிப்பகம், அச்சு, பக்கம், தாள்	இல்லை
விலங்கு	மிருகம்	காடு, சிங்கம், புலி, தோல், யானை, தந்தம், நாய், பூனை	கிழமைவேற்றுமை உடனிகழ்ச்சி வேற்றுமை
விலங்கு	(கை) விலங்கு	குற்றவாளி, கோர்ட், போலீஸ், கைதி, விசாரணை, தண்டனை	இல்லை

சிறப்பான வேற்றுமை உருபு இல்லாதது 'இல்லை' எனக் குறிக்கப்பட்டுள்ளது.

5.2.2.3 பரிசோதனைக் கட்டம்

இப்போது ஒழுங்குமுறை பயன்பத்துபவர் பயன்படுத்துவதற்குத் தயாராக உள்ளது. இதில் ஒருவர் பொருண்மை மயக்கச் சொற்கள் உள்ள தமிழ்ப் பனுவலை உள்ளீடாகக் கொடுக்க இயலும் மற்றும் முழுவதும் அல்லது பகுதி பொருண்மை மயக்கம் நீக்கப்பட்ட பனுவலை வெளியீடாகப் பெற இயலும். பரிசோதனை கட்டத்தில் அர்த்தம்-சேர்ந்துவருகை வேற்றுமை குறியீட்டு அகராதி பார்க்கப்படாத நேர்வுகளைப் பொருண்மை மயக்கம் நீக்குவதற்குப் பயன்படுத்தப்படுகின்றது.

பரிசோதனைக் கட்டத்தில் கண்டுபிடிக்கப்படும் சேர்ந்துவருவனவும் வேற்றுமைக் குறியீடுகளும் தரப்பட்ட பனுவலில் உள்ள பொருண்மை மயக்கங்களை நீக்கப் பயன்படுத்தப்படுகின்றது. இருப்பினும் வழிமுறை வரைவு தரப்பட்ட பனுவலில் தோன்றும் பொருண்மை மயக்கமுள்ள சொற்களுக்கும் முன்னரே பயிற்சிசெய்யப்பட்டிருக்கவேண்டும்.


பார்க்காத பனுவலில் நேர்வுகள்
நிறுத்தல் சொற்கள் நீக்கி மற்றும் பொருண்மை மயக்கச் சொற்கள் கண்டுபிடிப்பி
உருபனியல் பகுப்பாய்வி
சூழல் சொற்கள் பிரித்தல்
சூழல் சொற்களைப் பட்டியலிடப்பட்டுள்ள சேர்ந்துவருவனவற்று டன் ஒப்பிடுக
அர்த்தம்-சேர்ந்துவருவன வேற்றுமைக்குறியீடு அகரதி
வேற்றுமைக் குறியீடு பிரித்தெடுப்பு
பொருண்மைமயக்கம் நீக்கப்பட்ட நேர்வுகள்
அகராதியில் பட்டியலிடப்பட்டுள்ள வேற்றுமைக் குறியீடுகளுடன் ஒப்பிடுக
பொருத்தம் காணப்பட்டது தீர்மானிக்கப்பட்டது
பொருத்தம் காணப்படவில்லை, தீர்மானிக்கப்படவில்லை
பொருண்மை மயக்கம் நீக்கபடாத நேர்வு
முழுவது / பகுதி பொருண்மை மயக்கம் நீக்கப்பட்ட பனுவல்

5.3 கருத்துரு வரைபட அகராதி அடிப்படையில் சொற்பொருண்மை மயக்க நீக்கம்

இவ்வாய்வில் சொற்பொருண்மை மயக்கநீக்கத்திற்கு வேண்டி கருத்துரு வரைபட அகராதி/ கருத்துரு அகராதி உருவாக்கப்பட்டுப் பயன்படுத்தப்பட்டுகின்றது. கருத்துரு வரைபட அகராதியின் உருவக்கம் பற்றியும் அதைப் பயன்படுத்திச் சொற்பொருண்மை மயக்க நீக்கம் எவ்வாறு மேற்கொள்ளப்படுகின்றது என்றும் இங்கு விரிவாக விளக்கபட்டுள்ளது. இதற்கு அடிப்படையாக அமையும் சொற்பொருண்மை ஆய்வு (lexical semantics study) பற்றி முன்னர் விளக்கப்பட்டுள்ளது. இது அறிவு அடிப்படையிலான நெறிமுறையை/அணுகுமுறையைச் சாரும். இந்நெறிமுறையில் கருத்துரு வரைபட அகராதி சொற்களின் பொருண்மை மயக்கநீக்கத்திற்குப் பயன்படுத்தப்படுகின்றது. இது ஒரு மாதிரி (model) நிலையிலேயே அமைந்துள்ளது. பொருண்மை மயக்கம் தருகின்ற சில சொற்கள் தேர்ந்தெடுக்கப்பட்டுத் தரவுத்தொகுதியிலிருந்து அச்சொற்கள் வரும் வாக்கியங்கள் பிரித்தெடுக்கப்பட்டு பொருண்மை மயக்கநீக்கம் செய்ய முயற்சி எடுக்கப்பட்டுள்ளது. இதற்கு என்று அமிர்தாப் பல்கலைக்கழகத்தின் தரவுத்தொகுதி பயன்படுத்தப்பட்டுள்ளது முழுச் சொற்றொகைக்கும் தற்போது இதைப் பயன்படுத்த இயலாது. இருப்பினும் ஒரு முழுமையான கருத்துரு வரைபட அகராதி உருவாக்கப்பட்டால் இது ஓரளவுக்குச் சாத்தியமாகும்.

இது இருகட்டங்களைக் கொண்டது: (1) கற்றல் கட்டம் (2) பரிசோதனைக் கட்டம். முதல் கட்டத்தில் கருத்துரு வரைபட அகராதியைப் பயன்படுத்தி தரவுத்தொகுதியில் வரும் பொருண்மை மயக்கமுள்ள சொற்கள் வேறுபட்ட கருத்துருக்களுக்காக (அதாவது பொருண்மைகளுக்காக/அர்த்தங்களுக்காக) அடையாளப் படுத்தப்படும். இரண்டாம் கட்டத்தில் அடையாளப்படுத்தப்படாத தரவுத்தொகுதியில் உள்ள செற்களின் பொருண்மை மயக்கங்களை நீக்கப் பரிசோதனை செய்யபடும்.

5.3.1 கருத்துரு வரைபடம்

செயற்கை அறிவு வெளிப்படுத்தத்திற்குப் பயன்படுத்தப்படும் கருத்துரு சார்ந்த அமைப்பு அடிப்படையிலான கோட்பாடு சௌவா (Sowa, 1984) என்பவரால் விளக்கப்பட்டுள்ளது. இது மூளையிலும் இயந்திரத்திலும் (கணிப்பொறியிலும்) தகவலை ஆய்வதற்கான சாத்தியம் அடிப்படையில் அமைந்தது. கருத்துரு வரைபடம் நபர்களுக்கும் (persons) பொருளுக்கும் (things) நிகழ்வுகளுக்கும் (events) இடையிலுள்ள உறவுகளைக் கூறும் தர்க்க வடிவங்களாகும். சாம்ஸ்கி

மாற்றிலக்கணம் அறிதிறனின் ஒரு அருவக் கோட்பாடு என்றும் அது செயல் திறனின் கோட்பாடாகக் கருதப்படக்கூடாது என்பார். ஆனால் செயற்கை அறிதிறனுக்கு வழியமைப்பாளர்கள் மக்களுக்கும் கணிப்பொறிக்கும் இடையில் நடக்கும் கருத்துப்பரிமாற்றத்திற்கு உதவிபுரியும் செயல்திறன் கோட்பாட்டை வேண்டுவர். செயற்கை அறிதிறனில் கருத்துரு வரைபடம் பரவலாகப் பயன்படுத்தப்படுகின்றது. பின்வரும் கருத்துரு வரைபடம் எடுத்துக்காட்டாகத் தரப்பட்டுள்ளது.

பையன் நாயைக் கடித்தான்.

பின்வரும் கருத்துரு வரைபடம் இந்த வாக்கியத்தின் அடிப்படைப் பொருளை உருப்படுத்தம் செய்யும்.

[பையன்]← (செயலி) ←[கடி] →(செ.பொ) →[நாய்]

பெட்டிகள் கருத்துருக்கள் (செவ்வக அடைப்புக் குறிகளுக்குள் தரப்பட்டுள்ளது) என்றும் வட்டங்கள் (பிறை அடைப்புக்குறிக்குள் தரப்பட்டுள்ளது) கருத்துரு உறவுகள் என்றும் அழைக்கப்படும். கருத்துரு எந்த ஒரு மொழியில் வரும் எந்த இருப்புப் பொருளையும் (அதாவது செயலையோ நிலையையோ) உருப்படுத்தம் செய்யும். கருத்துரு உறவுகள் ஒவ்வொரு இருப்புப் பொருளும் பங்காற்றுகிற பங்களிப்பைக் குறிக்கும். மேற்கண்ட எடுத்துக்காட்டில் [பையன்] என்பது [கடி] என்ற செயலின் (செயலி)-ஆகச் செயல்படும் கருத்துருவை உருப்படுத்தம் செய்யும். [பையன்] என்ற கருத்துரு [கடி] என்பதன் (செயப்படுபொருள்)-ஆகப் பங்கெடுக்கும் இருப்புப் பொருளை உருப்படுத்தம் செய்யும். கருத்துரு வரைபடத்தார் எல்லா பொருண்மை குறித்த செய்திகளையும் ஒரு அடி அமைப்பில் இருப்பதாகக் கொள்ளாமல் ஆறு வேறுபட்ட வகையிலான செய்திகளிலிருந்து வாக்கியங்கள் ஆக்கப்படுவதாகக் கொள்கின்றனர். அவையாவன: இறந்தகாலமும் உள்ளது கிளத்தல் வினை நோக்கும் (indicative mood) செயல் குறிப்பிட்ட நேரத்திற்கு முன்னர் நிகழ்ந்ததைக் குறிப்பிடும். *பையன்* ஒரு சூழலில் ஒரு குறிப்பிட்ட நபரின் முன்வரு கிளவியாக வருகிறது என்பது முற்கோளாகும். *அவன்* சிறிது முன்னர் குறிப்பிடப்பட்டிருக்கலாம் அல்லது அந்த காட்சியில் இருக்கலாம். *நாய்* கடிப்பதைப் பற்றிய புதிய செய்தி கவனக்குவிப்பாகும். பையன் முன்னர் கூறப்பட்ட ஒருவரைக் குறிப்பதன் காரணமாக ஒரு உடனிலைக் குறிப்பு முன்னர் வரும் பையனை *பையன்* என்ற கருத்துருவுடன் தொடர்புபடுத்துகின்றது. ஆறாவதான உணர்ச்சி உணர்பொருள்கள் வாக்கியத்தில் உள்ள

வார்த்தைகளாலோ வரைபடத்திலுள்ள கருத்துருக்களாலோ நேரடியாக வெளிப்படுத்தப்படவில்லை. ஆனால் அது மக்களுக்குள் வேறுபடும் முந்தைய அனுபவங்களின் சேர்க்கைகளால் நிறுவப்பெறும்.

5.3.1.1 வாய்பாட்டுக் கருத்துரு வரைபடம் (Canonical Graph)

ஒரு கருத்துரு வரைபடம் கருத்துரு கணுக்கள் மற்றும் உறவுக் கணுக்கள் இவற்றின் ஒருங்கிணைப்பாகும். இதில் ஒவ்வொரு கருத்துரு உறவின் ஒவ்வொரு வில்லும் (இங்கு அம்புக்குறி) ஒரு கருத்துருவுடன் தொடர்புபடுத்தப்பட்டுள்ளது. ஆனால் எல்லா ஒருங்கிணைப்பும் அர்த்தம் தராது. சில பின்வருமாறு பொருளற்ற ஒருங்கிணைப்பைக் காட்டும்.

[தூங்கு]→ (செயலி) →[கருத்து]←(நிறம்) ←[பச்சை]

'பச்சைக் கருத்து தூங்குகிறது'

வெளியுலகில் உண்மையான அல்லது சாத்தியமான சூழல்களை உருப்படுத்தம் செய்யும் பொருள் தரும் கருத்துரு வரைபடத்தை வேறுபடுத்த சில வரைபடங்கள் வாய்ப்பாட்டு வரைபடங்கள் எனக் கூறப்படுகின்றன.

[சிறுமி] ← (செயலி) ← [உண்] →(முறை) → [விரைவாக]

[நபர்: ராமன்] ← (செயலி) ←[சாப்பிடு] → (செபொ) → [கடலை]

'ராமன் கடலை சாப்பிடுகிறான்'

5.2.1.2. பொதுமையாகமும் சிறப்பாக்கமும் (generalization and specialization)

வாய்ப்பாட்டு வடிவமாக்க விதிகள் யாவும் சிறப்பாக்க விதிகளாகும்.

ஒரு பொதுவான பொதுமையாக்கமும் ஒரு பொதுவான சிறப்பாக்கமும்.:

[நபர்]←(செயலி)←[உண்]

[சிறுமி]←(செயலி)←[உண்]→(முறை)→[விரைவாக]

[நபர்:ராணி] ←(செயலி)←[சாப்பிடு]→(செபொ)→[கடலை]

[சிறுமி:ராணி]←(செயலி)←[உண்]→(முறை)→[விரைவாக]

↓

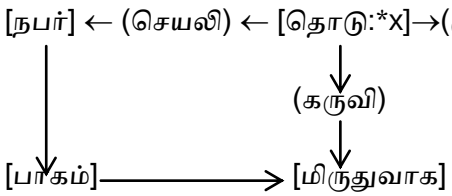
(செபொ) → [கடலை]

5.2.1.3 சுருக்கமும் வரையறை விளக்கமும் (Abstraction and definition)

வரையறை விளக்கம் ஒரு வகையை (type) இரு வேறுபட்ட வழிகளில் குறிப்பிடலாம்: வகையின் தேவையான மற்றும் போதுமான கட்டுப்பாடுகளைத் தருவது; அல்லது சில

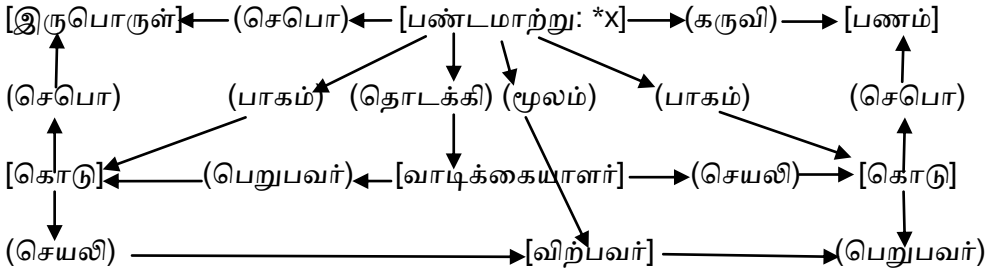
எடுத்துக்காட்டுகளைத் தந்து அவற்றுடன் ஒற்றுமையுள்ள எல்லாம் இவ்வகையைச் சாரும் என்பது. முதலாவது முறை அரிஸ்டாட்டலின் பொதுக்கூறு (genus) மற்றும் சிறப்புக்கூறு (differentiae) என்பதால் வரையறை விளக்கம் செய்வதிலிருந்து ஆக்கப்பட்டதாகும். பொதுக்கூறு மற்றும் சிறப்புக்கூறு இவற்றால் செய்யப்படும் வரையறை விளக்கம் தர்க்க அடிப்படையில் கையாளுவதற்கு எளிதாகும். எடுத்துக்காட்டுகளாலும் மூல முன்மாதிரிகளாலும் செய்யப்படும் வரையறை விளக்கம் இயற்கை மொழிகளையும் உண்மை உலகில் அதன் பயன்பாடுகளையும் கையாளுவதற்கு முக்கியமானதாகும். கருத்துரு வரைபடங்கள் பொதுக்கூறுகளாலும் சிறப்புக்கூறுகளாலும் மற்றும் திட்டவரைவுகளாலும் (schemata) மூல முன்மாதிரிகளாலும் (prototype) வரையறை விளக்கம் செய்வதை ஆதரிக்கின்றது. இந்த இரு முறைகளும் வடிவ அளவீடுகளாகக் குறிப்பிடும் ஒன்றோ அதற்குக் கூடுதலான கருத்துருக்கள் கொண்ட கருத்துரு வரைபடங்களான சுருக்கங்களின் அடிப்படையில் ஆகும்.

முத்தமிடு என்பதன் வகை வரையறை விளக்கம் பின்வருமாறு அமையும்:



முத்தமிடு: 'ஒரு நபர் உதடுகளால் மிருதுவாகத் தொடுதல்'

வாங்கு என்பதன் வரையறை விளக்கம் பின்வருமாறு அமையும்:



5.2.1.4. ஒன்றுதிரட்டலும் தனிநிலைப் படுத்தலும் (aggregation and individuation)

ஒவ்வொடு பன்மைப் பெயரும் ஒரு குழுமத்தை உருப்படுத்தம் செய்யும். ஒரு குழுமத்தை உருப்படுத்தம் செய்ய ஒரு கருத்து மீசை அடைப்புக்குறிக்குள் மூடப்பட்ட ஒரு குழுமப்பெயர்களையோ தனிநிலை குறியீடுட்களையோ கொண்டிருக்கும்: [நாய்:{ஜிம்மி, பப்பி, ரோசி}].

மூன்று வேறுபட்ட பயன்பாடுகளைப் பன்மைப் பெயர்த்தொடர்களுக்கு வேறுபடுத்தலாம்:

தொகை (collective): ஒரு குழுமத்தின் எல்லாத் தனிமங்களும் சேர்ந்து ஏதோ ஒரு உறவில் பங்களிப்புச் செய்யும். எ.கா. ராஜாவும் ராணியும் அந்த வீட்டின் சொந்தக்காரர்கள்.

பகிர்வு (distributive): ஒரு குழுமத்தின் ஒவ்வொரு தனிமங்களும் ஏதோ ஒரு உறவைத் திருப்திசெய்யும். எ.கா. ராஜாவும் ராணியும் சிரிக்கிறார்கள்.

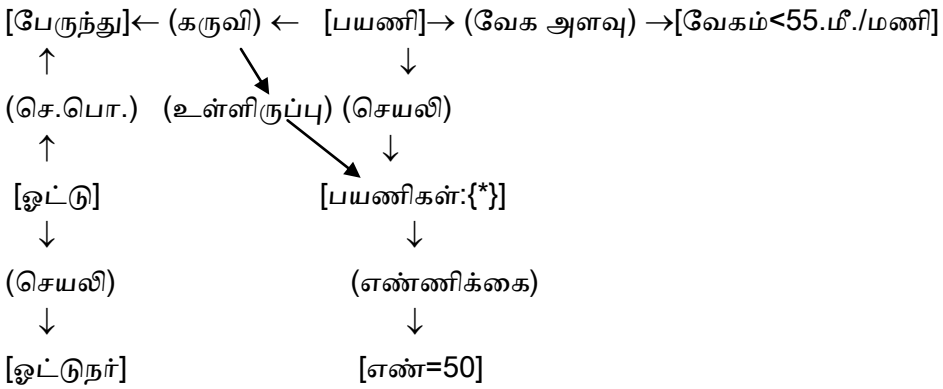
முறையே: ஒரு நிரல்படுத்தப்பட்ட வரிசையின் ஒவ்வொரு தனிமமும் மற்றொரு வரிசையின் பொருத்தமான தனிமத்துடன் ஒரு குறிப்பிட்ட உறவு கொண்டிருக்கும்.

கருத்துரு வரைபடத்தில் மீசை அடைப்புக்குறிகள் ஒரு தொகைப் பொருள்கோளைக் காட்டும்: {ராஜா, ராணி, கண்ணன்} (அதாவது ராஜாவும் ராணியும் கண்ணனும்). ஒரு குழுமத்தின் முன்னால் வரும் 'பகிர்' பகிர்ந்துகொள்ளப்படும் பொருள்கோளைக் காட்டும்: பகிர் {ராஜா, ராணி, கண்ணன்}. கோண அடைப்புக்குறிகளுக்குள் (<, >) வரும் ஒரு குழுமத்தின் முன் வரும் 'முறையே' என்ற குறிப்பு முறையே வரும் பொருள்கோளைக் காட்டும்: முறையே <ராஜா, ராணி, கண்ணன்>. பிரிநிலைக் குழுமத்தில் (disrtibutive set) வரும் தனிமங்கள் செங்குத்துகோடால் பிரிக்கப்படும்: {ராஜா| ராணி | கண்ணன்} (அதாவது ராஜா அல்லது ராணி அல்லது கண்ணன்).

5.2.1.5 திட்டவரைவும் முன்மாதிரியும் (Schemata and Prototype)

மனித மாதிரியான புரிந்துக்கொள்ளலை உருப்படுத்தம் செய்யும் அடிப்படை அமைப்பைத் திட்டவரைவு என்பர். திட்டவரைவு கருத்துரு வரைபடத்தின் மூன்றாம் நிலை கலவையாகும். இடுகுறியான கருத்துரு வரைபடங்கள் அனுமதிக்க இயலும் இணைப்புகளுக்கு எந்தக் கட்டுப்பாட்டையும் விதிக்காது. வாய்ப்பாட்டு கருத்துரு வரைபடங்கள் (canonical form) தேர்வுக் கட்டுப்பாடுகளை விதிக்கும். இவை மொழியியலில் வேற்றுமை சட்டகத்துடனும் தத்துவத்தில் வகைப்பாட்டுக் கட்டுப்பாடுகளுடன் ஒப்புமைப்படுத்தத் தக்கது. திட்டவரைவு (schemata) உண்மை உலகிலுள்ள இருப்புப்பொருள்கள், அடைகள், நிகழ்வுகள் இவற்றின் கூட்டத்தைப் பற்றிய குறிப்பிட்ட பொருட்புல அறிவை உட்படுத்தும்.

பேருந்து என்பதற்கான திட்டவரைவு கீழே தரப்பட்டுள்ளது.



முன்மாதிரி என்பது ஒரு மாதிரி நிகழ்வு. ஒரு குறிப்பிட்ட நபரை/ஒன்றை விளக்காமல் ஒரு மாதிரி சராசரி நபரை/ஒன்றை விளக்கும். எடுத்துக்காட்டாக ஒரு யானையின் திட்டவரைவு ஒரு எல்லைக்குட்பட்ட பண்புகளையோ பழக்கவழக்கங்கள் மற்றும் இருப்பிடங்களையோ குறிப்பிடும். முன்மாதிரி (prototype) அம்மாதிரியான திட்டவரைவுகளை ஒருங்கிணைத்தும் கட்டுப்படுத்தியும் ஒரு மாதிரி யானையை விளக்கும். பின்வருவது யானையின் முன்மாதிரியாகும்.

(யானை:*x)

(பண்பு)→(உயரம்:3.3 மீ)

(பண்பு→ (எடை: 5400 கி.கி.)

(நிறம்) → (கறுப்பு)

(பாகம்) → (மூக்கு)

(அடை)

(பிடிகை) → (துதிக்கை)

(பாகம்) → (காது: *})

(எண்ணிக்கை)→ (எண்:2)

(அடை)→ (முறம் போன்ற)

(பாகம்) →(கொம்பு:({ *})-

(எண்ணிக்கை)→ (எண்:2)

(செய்பொருள்)→ (தந்தம்)

(பாகம்) → (கால் { *})-

(எண்ணிக்கை)→ (எண்:4)

(நிலை) → (வாழ்)-

(வாழ்விடம்) → (கண்டம்: {ஆப்பிரிக்கா)

(வாழ்நாள்) → (காலம்: 50 ஆண்டுகள்)

கருத்துரு வரைபடம் அடிப்படையில் தமிழ் வாக்கியங்களை ஆயும் இவ்வாய்வு ஒரு தொடக்க நிலையில் தான் உள்ளது. சௌவா கருத்துருக்களுக்கும் கருத்துருக்களுக்கம் இடையிலுள்ள பல உறவுகளைப் பட்டியலிடு விளக்குகின்றார். அவையெல்லாம் தமிழுக்கும் பொருந்துமாறு இவ்வாய்வில் மாற்றியமைக்கப்பட்டுள்ளது.

5.2.2 கருத்துரு வரைபட அகராதி உருவாக்கத்திறகான மூலவளங்கள்

இவ்வாய்வேட்டில் சொற்களஞ்சியம் மற்றும் அகராதி அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் முயற்சிக்கப்பட்டுள்ளது. இராசேந்திரன் (இராசேந்திரன் 2001) அவர்களின் தற்காலத் தமிழ்ச்சொற்களஞ்சியமும் கிரியாவின் தற்காலத் தமிழ் அகராதியும் இங்கு பயன்படுத்தப்பட்டுள்ளது.

5.2.2.1. இராசேந்திரனின் மின்சொற்களங்சியம்

இராசேந்திரானில் (2006) தரப்பட்ட தமிழ்சொற்றொகையின் பொருண்மை அமைப்பு (அதாவது படிநிலைப் பாகுபாட்டு ஒழுங்கமைப்பு) வழியமைப்புமுறைகளால் (programs) மின்சொற்களஞ்சியத்திற்கான தரவு அடிப்படையாக மாற்றப்பட்டுள்ளது. இதிலிருந்துதான் பயன்படுத்துபவர் கேட்கும் கேள்விகளுக்கான தகவல் தரப்படும். சொற்களஞ்சியச் சொற்றொகை பின்வருமாறு பொருண்மைப் புலங்களாகப்பகுக்கப்பட்டுள்ளது (semantic domains/fieds), ஒவ்வொரு பொருட்புலனுக்கும் படிநிலை அடிப்படையிலும் முன்நிலை அடிப்படையிலும் எண்கள் தரப்பட்டுள்ளன. இவ்வெண்களின் வரிசை சொற்றொகைப் பாகுபாட்டின் பொருண்மை அமைப்பை விளக்குவதாய் அமையும். தரவு அடிப்படை புதிய சொற்களைச் சேர்க்கவும் புதிய பாகுபாடுகளை உருவாக்கவும் வேண்டி நெகிழ்வுள்ளதாய் அமைக்கப்பட்டுள்ளது. இரண்டாம் நிலையில் தரப்படும் புதிய செய்திகள் தற்போதைய தரவு அடிப்படையுடன் உறவுபடுத்தப்படும். பின்வரும் சொற்றொகைப் பாகுபாடு மின்சொற்களஞ்சியத்தின் முதன்நிலை தரவு அடிப்படையை உருப்படுத்தம் செய்யும்.

உளமொழியல் கோட்பாடுகளை எந்த அளவிற்குப் பயன்படுத்தி மின்சொற்களஞ்சியம் அமைக்கலாம் என்பது ஆய்வுக்குரியது. பெரும்பாலும் இத்தகைய ஆய்வுகள் பெயர்களைச் சுற்றி வருகின்றனவே தவிர பிற சொல்வகைப்பாட்டைச் சார்ந்த வினைகள், பெயரடைகள் பற்றிய ஆய்வுகள் அதிகம் இல்லை. இச்சொற்களஞ்சியம் ஒரு அகராதியை அடிப்படையாகக் கொண்டு அதில் உள்ள சொற்களைப் பொருண்மை அடிப்படையிலும் அவற்றின் பொருண்மை உறவு அடிப்படையிலும் தொடர்புபடுத்தும் நோக்கில் அமைந்துள்ளது. சொற்களஞ்சியத்தைப் பொறுத்தமட்டில் பெயர், வினை, பெயரடை, வினையடை ஆகிய சொல்வகைகளைச் சார்ந்த சொற்கள் தரவுகளாகத் தேர்ந்தெடுக்கப்பட்டுள்ளன. இலக்கணச் சொற்கள் தொடர்பன்கள் என்ற நிலையில் சேர்த்துக் கொள்ளப்பட்டுள்ளன. தேர்ந்தெடுக்கப்பட்ட சொற்கள் யாவும் பொருண்மைக்கூறுகள் அடிப்படையில் வகைபடுத்தப்பட்டு வகைப்பாட்டு எண்களுடன் பட்டியலிடப்பட்டு தரவு அடிப்படையாகச் சேகரிக்கப்பட்டுள்ளன.

மின்சொற்களஞ்சியத்தின் மிகப் பரந்த நோக்கம் என்னவென்றால் சொற் தகவல்களை சொல்வடிவுகள் அடிப்படையில் அல்லாமல் சொற்பொருண்மை அடிப்படையில் ஒழுங்குபடுத்துவதாகும். இந்த அடிப்படையில் மின்சொற்களஞ்சியம் அகராதியிலிருந்தும் வேறுபடும். ஆங்கிலத்திற்கு உருவாக்கப்பட்ட ரொஜஸ்ட் தெசாரஸ் போன்ற சொற்களஞ்சியங்களும் தமிழுக்கு இராசேந்திரனால் (இராசேந்திரன், 2001) உருவாக்கப்பட்ட தற்காலத் தமிழ்ச் சொற்களஞ்சியமும் தமிழ் மின்சொற்களஞ்சிய உருவாக்கத்திற்கு முன்னோடியாய் அமைகின்றன. மேற்கூறிய சொற்களஞ்சியங்கள் புத்தக வடிவில் அமைந்தவை. அகரவரிசையில் அமைந்த சொற்களஞ்சியங்களின் சிக்கல் என்னவென்றால் அவற்றில் தேவைக்கு அதிகமான/மிகையான பதிவுகள் காணப் பெறும். இரு சொற்கள் ஒரு பொருள்பன்மொழிகள் என்றால் கூட அகரவரிசைக்காக இருமுறை பதிவு பெறும். தலைப்பு ஒழுங்குமுறையில் அமைந்த சொற்களஞ்சியங்களின் சிக்கல் என்னவென்றால் இரண்டுமுறை நோக்கீடு செய்துதான் தேவையான தகவல்களைப் பெரும்பாலும் பெற இயலும். முதல்முறை அகரவரிசையில் அமைந்த சொல்லடைவு மூலம் குறிப்பிட்ட சொல் எந்தப் பொருண்மைத் தலைப்பில் வருகிறது என்பதைத் தெரிந்து கொண்டு இரண்டாவது முறை நோக்கீடு செய்துதான் வேண்டிய தகவலைப் பெறமுடிகிறது. இதனால் தேடும் நேரம் இரண்டு மடங்காக உயரும். கணிப்பொறியின் வழி தேடலை விரைவாகச் செய்ய இயலும்.

சொல்வடிவுகளுக்கும் சொல் பொருண்மைகளுக்கும் உள்ள பொருத்தம் பல. சில வடிவுகள் வேறுபட்ட பொருண்மைகளைத் தரும்: சில பொருண்மைகளைப் பல வேறுபட்ட வடிவங்களில் வெளிப்படுத்தலாம். பல்பொருள் ஒருமொழியம் (polysemy), ஒருபொருள் பன்மொழியம் (synonymy) என்பன அகராதியியலின் இரு கடினமான சிக்கல்களாகும். இவற்றைப் பொருத்துவதின் துணைநிலைகளாகக் கொள்ளலாம். உள அகராதியை (mental dictionary) அணுகுவதன் வாயிலாக பல்பொருள் ஒருமொழியமும் ஒருபொருள் பன்மொழியமும் சிக்கல்களாகத் தெரிகின்றன. ஒரு வடிவத்தைத் தெரிந்து கொள்ளும் கேட்பவரோ படிப்பரோ இந்தப் பல்பொருள் ஒருமொழியத்தை எதிர்கொள்ள வேண்டி வரும்; பொருளை வெளிப்படுத்த விரும்பும் பேசுபவர் அல்லது எழுதுபவர் ஒருபொருள்பன்மொழிகளுக்கிடையில் ஒன்றை தேர்ந்தெடுக்க வேண்டிவரும்.

மின்சொற்களஞ்சியம் பொருண்மை உறவுகளுக்கும் சொல் உறவுகளுக்கும் இடையே உள்ள வேறுபாட்டை உணர்த்துகின்றது. பொருண்மைகளுக்கு இடையே உள்ள வேறுபாட்டை உணர்த்துகின்றது. பொருண்மைகளுக்கு இடையே உள்ள பொருண்மை உறவுகளுக்கு முக்கியத்துவம் கொடுக்கப்பட்டாலும், சொற்களுக்கு இடையே உள்ள உறவுகளும் உட்படுத்தப்பட்டுள்ளன.

இராசேந்திரன் அவர்களின் மின் சொற்களஞ்சியம் லையான்ஸ் அவர்களின் அமைப்புப் பொருண்மையியல் (structural linguistics) கோட்பாடு அடிப்படையிலும் நைடாவின் பொருண்மைகூறுகள் ஆய்வு (componential analysis of meaning) அடிப்படையிலும் அமையும். நைடா (1975a) அண்மைப்படுதல் (contiguity), மேலுறல் (overlapping), உள்ளடங்கல் (inclusion), துணைநிலைப்படுதல் (complementation) ஆகிய உறவுகளால் சொற்றொகையைப் பாகுபடுத்துகிறார். நைடா அண்மைப்படுதலை அடிப்படை உறவாகக் கருதுகிறார். நைடா குறிப்புப் பொருளின் (referential meaning) வேறுபடல் பொருள் கூறு ஆய்வை (componential analysis of meaning) ஒரு மொழியின் சொற்றொகையின் பொருண்மை அமைப்பைக் காண்பதற்குப் பயன்படுத்துகிறார். சொல்லுக்கும் அதன் குறிப்புக்கும் உள்ள தொடர்பு தான் குறிப்புப் பொருளாகும். நைடா (1975:152) பொருள் உணர்த்தும் பலவகைச் சொற்சேர்க்கைகள் ஆய்வு நம்மை நான்கு அடிப்படைப் பொருண்மைக் குறியீடுகளின் வகுப்புகளை இனங்கண்டு கொள்ளத் தூண்டும்.

1. பருப்பொருள்கள்

   எ.கா. குதிரை, மரம், இலை.

2. நிகழ்வுகள் (வினைகள்)

   எ.கா. ஓடுதல், நடத்தல், வெட்டுதல்.

3. பெயரடைகள் (பருப்பொருள்களின் அருவங்கள்)

   எ.கா. சிவப்பு, நீலம், சின்னது.

4. வினையடைகள் (நிகழ்வுகளின் அருவங்கள்)

   எ.கா. விரைவு, மெல்ல.

5. பல பருப்பொருள், நிகழ்வுகள், அருவங்கள்இவற்றைத் தொடர்புபடுத்த உதவும் தொடர்பான்கள்.

   எ.கா வீட்டிலுள்ள மனிதர்கள்

   பறக்கும் மனிதர்கள்

   ஓட்டமும் குதியும்

   பொருளின் அழகு

இந்த நான்கு பொருண்மை அமைப்புகளும் எல்லா மொழிகளிலும் காணப்படுகிறது. இவை எல்லா வகையான சொல்லுருக்களுக்கிடையேயுள்ள பொருள் அடிப்படையிலான தொடர்பை ஆய்வதற்கு அடிப்படையாக அமைகிறது. நைடாவின் பொருண்மையமைப்பு/பொருட்புலப்பாகுபாடு பின்வருமாறு பட்டியலிடப்பட்டுள்ளது.

1. பருப்பொருள்

எ. விலங்குகளல்லாதவை

1. இயற்கையானவை

   எ. பூகோளம் தொடர்புடையவை

   பி. இயற்கைப் பொருட்கள்

   சி. தாவரங்கள் மற்றும் தாவரப்பொருட்கள்

2. உருவாக்கிய அல்லது கட்டப்பட்ட பொருட்கள்

   எ. கலைப் பொருட்கள்

   பி. பதன் செய்யப்பட்ட பொருட்கள்: உணவுகள் மட்டும் வாசனைப் பொருட்கள்

சி. கட்டிடங்கள்

பி. விலங்குப் பொருட்கள்

1.விலங்குகள், பறவைகள், பூச்சிசள்

2. மனிதர்கள்

3.இயற்கைக்கு அப்பாற்பட்ட சக்திகள் அல்லது உயிர்கள்

2. நிகழ்வுகள்

பௌதிக நிகழ்வுகள், உடற்கூறு தொடர்புடைய நிகழ்வுகள், புலனுணர்வுகள், உணர்ச்சிகள், அறிவு தொடர்புடைய நிகழ்வுகள், கருத்துப் பரிமாற்றம், சேர்க்கை, ஆளுகை, சலனம், தாக்கம், கைமாற்றம், பல்கூட்டான செயல்கள்

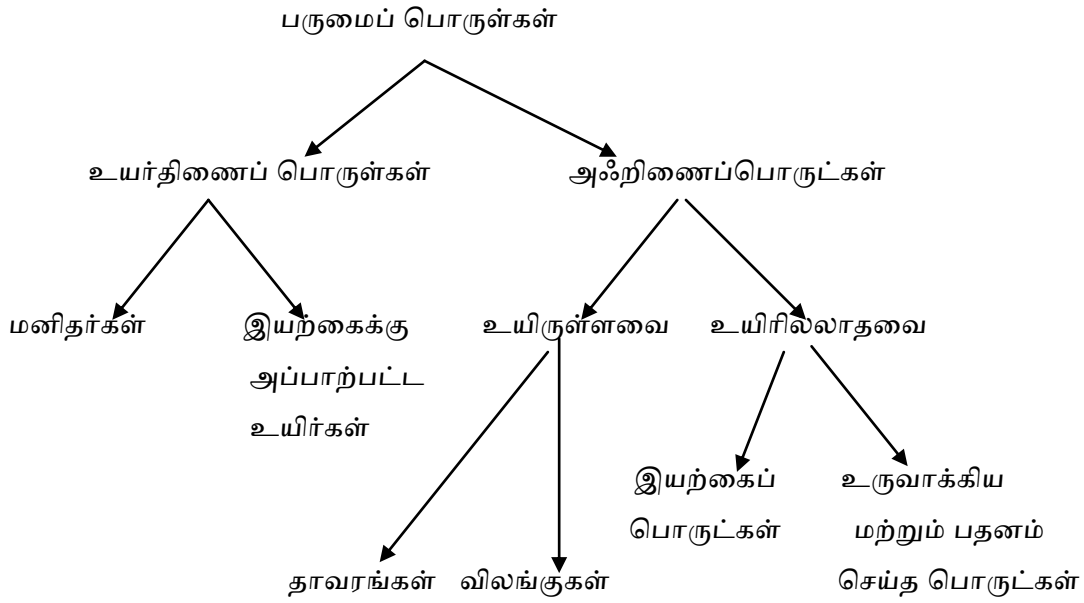
3. அருவங்கள்

காலம், தூரம், பருமை, விரைவு, சூடு, நிறம், எண், நிலை, மதப்பண்பு, கவர்ச்சி, வயது, உண்மை-பொய், நல்லது-கெட்டது, பலம், ஆரோக்கிய நிலை

4. தொடர்பன்கள்

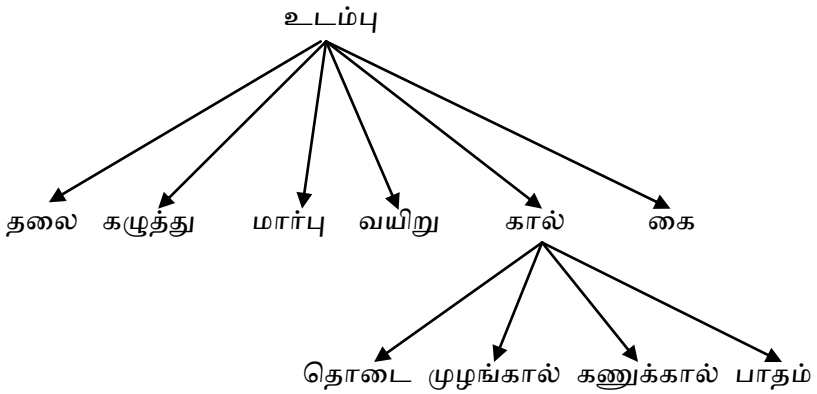
இடம் தொடர்பானவை, காலம் தொடர்பானவை, சுட்டல், தர்க்கம் தொடர்பானவை

பெயர்கள் எல்லாவற்றையும் ஒரே படிநிலை அமைப்பில் தரவேண்டி படிநிலைக் கொள்கையை நீட்சி செய்யலாம். சொல்வலை பெயர்களைத் தன்மையான தொடக்கிகள் (unique beginners) அமையப் படிநிலை அமைப்புகளாகப் பகுத்துள்ளது. தனித்தன்மையான தொடக்கிகள் சொற்பொருண்மையியல் பொருண்மைக்கூறாய்வின் பொருண்மைக் கூறுகளுடன் ஓரளவுக்குப் பொருந்தும்.



சினை-முழு உறவு/சினைமொழிய-முழுமொழிய உறவு

சொற்களைப் படிநிலை அமைப்பில் தருவதில் சினை-முழு உறவு முக்கியப் பங்கு வகிக்கின்றது.



சினைமொழிய-முழுமொழிய சினை-முழுஉறவு பின்வரும் துணை வகைகளை உள்ளடக்கும்:

1. முழுப் பொருளுக்கும் அதன் உறுப்புகளுக்கும் இடையிலான உறவை வெளிப்படுத்துவன (பாகம், எ.கா. கை – விரல்)

2. முழுமைக்கும் அதிலிருந்து பிரிந்த பகுதிக்கும் உள்ள உறவை வெளிப்படுத்துவன (பகுதி, எ.கா. துண்டு – உலோகம்)
3. இடங்களுக்கும் விரிந்த இடங்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (இடம், எ.கா. பாலைவனச்சோலை: பாலைவனம்)
4. குழுமத்திற்கும் அதன் அங்கத்தினர்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (எ.கா. மந்தை - ஆடு)
5. பொருள்களுக்கும் அது உருவாக்கப்பட்ட உருப்பொருள்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (ஆன, எ.கா. புத்தகம் – காகிதம்)

பின்வரும் அட்டவணையில் பெயர்வலையில் கூறப்பட்டுள்ள பொருண்மை-சொல் உறவுகள் பட்டியலிடப்பட்டுள்ளன:

உறவுகள்	துணை வகைகள்	எடுத்துக்காட்டுகள்
ஒருபொருள் பன்மொழி		புத்தகம், நூல்
உள்ளடங்கு-உள்ளடக்கு		விலங்கு-பலூாட்டி
உள்ளடக்கு-உள்ளடங்கு		பசு-பாலூாட்டி
முழு-சினை	முழுமை-பாகங்கள்	மேசை-கால்
“	பாகம்-முழுமை	சக்கரம்-வண்டி
“	அங்கத்தினர்கள்-குழு	படைத்தலைவர்-படை
“	குழு-அங்கத்தினர்கள்	துறை-பேராசிரியர்
“	பகுதி-முழுமை	துளி-கண்ணீர்
“	இடம்-பரந்த இடம்	பாலைவனச்சோலை-பாலைவனம்
“	பொருள்-உருப்பொருள்	புத்தகம்-தாள்
ஈரிணை எதிர்நிலை	நிரலாக்கம் செய்யத்தக்கது	நல்லவன் – கெட்டவன்
“	துணை எதிர்நிலை	பகல் – இரவு
“	தனிப்பட்டவை	அஃறிணை– உயர்திணை
“	துருவ எதிரிநிலை	ஆண் – பெண்
“	பரஸ்பர சமூகப்பாத்திரங்கள்	மருத்துவர் – நோயாளி

“	சொந்தங்கள்	அம்மா – மகள்
“	கால உறவுகள்	காலை – மாலை
“	இட உறவுகள:செங்கோண	வடக்கு-கிழக்கு, மேற்கு
“	உறவுகள்	
“	இட உறவுகள்: நேர் எதிர்நிலை	வடக்கு –தெற்கு
“	பல்லிணை எதிர்நிலை: சங்கிலி	ஒன்று, இரண்டு, மூன்று
“	சுற்று	ஞாயிறு, திங்கள், செவ்வாய், புதன், வியாழன், வெள்ளி, சனி

மின்சொற்களஞ்சியத்தின் செயற்பாடு

மின்சொற்களஞ்சியத்தில் பயன்படுத்துபவர்க்கு என்று அமைக்கப்பட்டுள்ள இடைமுகத்தைப் (inteface) பயன்படுத்தி வேண்டிய செய்திகளைப் பெறலாம். கீழே தரப்பட்டுள்ளவை ஒரு சில பயன்பாடுகளாகும். இவை மின்சொற்களஞ்சியத்தின் முதல் நிலை அடிப்படையில் அமைந்தவை.

ஒருபொருள்பன்மொழிகளை அறிந்துகொள்ளுதல்

இடைமுகத்தில் தெரிந்த ஒரு சொல்லைத் தந்து அதனோடு நெருங்கிய தொடர்புள்ள சொற்களைப் பெறலாம். எடுத்துக்காட்டாக *நல்லவர்* என்பதை இடைமுகத்தில் உள்ளீடு செய்து அது தொடர்பான சொற்களைத் தரச் சொன்னால் *நல்லவர்* என்பதன் ஒருபொருள்பன்மொழிகள் ‘பண்புத்தொடர்புடைய பெயர்கள்’ என்ற தலைப்பின் கீழ்வரும் *நல்லவர்* என்ற துணைத் தலைப்பின் கீழ் பட்டியலிடப்படும்:

நல்லவர் (நல்லவன், நல்லவள், நல்லவள்), உத்தமர் (உத்தமன், உத்தமி), ஏந்தல்.

பகுதி ஒருபொருள்பன்மொழிகள் மற்றும் தொடர்புடைய சொற்களை அறிந்துகொள்ளுதல்

நல்லவர் என்பதுடன் தொடர்புடைய கூடுதல் சொற்களைத் தரச்சொன்னால் பின்வரும் சொற்களும் பட்டியலிடப்படும்.

பெரியவர் (பெரியவன்), பெரியமனிதர் (பெரியமனிதன்), பெரியோர், உயர்ந்தவர் (உயர்ந்தவன், உயர்ந்தவள்), சிறந்தோர் (சிறந்தோன், சிறந்தோள்), அடிகள், ஆன்றோர், மேன்மக்கள், உத்தமபுருஷர், மகாபுருஷர், மகான், மகாத்மா, மானஸ்தர் (மானஸ்தன், மானஸ்தி), மானி

எதிர்மறைப் பொருளில் வரும் சொற்களை அறிந்து கொள்ளுதல்

'பண்பு தொடர்புடைய பெயர்கள்' என்ற தலைப்பின் கீழ் *நல்லவர்* என்ற சொல்லுக்கு எதிரிடையான சொற்களையும் அறிந்துகொள்ளலாம்.

கெட்டவர் (கெட்டவன்:கெட்டவள்), கொடியவன்:கொடியவள்), கொடியோர் (கொடியோன்:கொடியோள்), தீயவர் (தீயவன்:தீயவள்), பொல்லாதவர் (பொல்லாதவன்: பொல்லாதவள்), குருரர் (குருரன்:குரூரி), வக்கிரர் (வக்கிரன்), கிராதகர் (கிராதகன்:கிராதகி), துஷ்டர் (துஷ்டன்:துஷடை), போக்கிரி, வம்பர் (வம்பன்:வம்பி), விஷமி, அக்கிரமி, அயோக்கியர் (அயோக்கியன்), கயவாளி, கசவாளி.

படிநிலை அமைப்பின் மூலம் உள்ளடங்கு-உள்ளடங்கு பொருண்மைத் தொடர்பு மற்றும் பொருண்மைக் கூறுகளை அறிந்து கொள்ளுதல்

நல்லவர் என்பது படிநிலை அமைப்பில் பின்வருமாறு அமைவதைப் பார்த்தோம்: பருமைப்பெயர்கள் > உயர்திணைப்பெயர்கள் > மனிதர்கள் > பண்புகுறித்த பெயர்கள் > நல்லவர் மேற்கண்ட படிநிலை அமைப்பிலிருந்து நல்லவர் மேற்கண்ட படிநிலை அமைப்பிலிருந்து *நல்லவர்* என்ற சொல் பண்புகுறித்த பெயர்கள் என்பதன் உள்ளடங்கு மொழியாகவும் பண்புகுறித்த பெயர்கள் யாவும் மனிதர்கள் என்பதன் உள்ளடங்குச் மொழிகளாகவும் மனிதர்கள் என்பதன் கீழ்வரும் சொற்கள் யாவும் உயர்திணைப்பெயர்கள் என்பதன் உள்ளடக்கு மொழிகளாகவும் உயர்திணைப்பெயர்கள் பருமைப்பெயர்களின் உள்ளடக்கு மொழிகளாகவும் வருவதை அறிந்துகொள்ளலாம். மட்டுமன்றி அவை உள்ளடக்கும் பொருண்மைப் பண்புக்கூறுகளையும் அறிந்து கொள்ளலாம்.

நல்லவர்

+ பருமைப்பெயர்

+ உயர்திணைப்பெயர்

+ மனிதப்பெயர்

+ பண்புகுறித்தபெயர்

சொற்கள் பொருண்மைக்கூறுகளால் ஒன்றுபடலையும் வேறுபடலையும் அறிந்து கொள்ளுதல்

நல்லவர் என்பதன் ஒருபொருள்பன்மொழியான *உத்தமர்* என்ற சொல்லுக்கும் *நல்லவர்* என்பதற்குத் தரப்பட்ட பொருண்மைக்கூறுகளைத் தரலாம்.

உத்தமர்

+ பருமைப்பெயர்

+ உயர்திணைப்பெயர்

+ மனிதப்பெயர்

+ பண்புகுறித்தபெயர்

எதிர்மறைப்பொருளில் வரும் *கெட்டவர்* என்ற சொல் *நல்லவர்* என்ற சொல்லிலிருந்து 'பண்பு' என்ற பொருண்மைக்கூறால் வேறுபடலையும் உணரலாம்.

கெட்டவர்

+ பருமைப்பெயர்

+ உயர்திணைப்பெயர்

+ மனிதப்பெயர்

+ பண்புகுறித்தபெயர்

இங்கும் இதற்கு முன்னர் கூறிய பத்திகளில் உள்ள செய்திகள் ஒரு அகராதிச் சொற்களஞ்சியம் (thesaurus with alphabetical entries) உருவாக்கப்பயன்படும்.

சொற்களின் விளக்கப்பொருண்மையை அறிந்துகொள்ளுதல்

சொற்கள் சொற்களஞ்சியத்தில் வரும் இடம் அதன் பொருபொருண்மையைத் தீர்மானிப்பதோடு அச்சொற்களின் பொருள் விவரணையையும் தருகின்றது. எடுத்துக்காட்டாக *நல்லவர்* என்ற சொல் பண்மைக்குறிக்கிற மனிதப் பெயர் என்ற பொருண்மை பெறும். *அறிஞன்* என்ற சொல்லுக்கு அறிவுத்தொடர்புடைய மனிதப் பெயர் என்ற விளக்கத்தைப் படிநிலை அமைப்பிலிருந்து பெறலாம். *கொடு* என்ற வினைக்கு ஒருவரின் கையிலிருந்து ஒரு பொருள் மற்றொருவர் கைக்குப் போவதைக் குறிக்கும் உடைமைமாற்ற வினை என்ற விளக்கம் பெறலாம். *திங்கள்* என்பது வாரத்தின் முதல்நாள் என்ற விளக்கத்தைப் பெறலாம். *கண்* தலையின் பகுதி என்ற விளக்கத்தைப் பெறலாம்.

இத்தகைய சொற்களஞ்சியம் நமது ஆய்வுக்கு மிகவும் பயனுள்ளதாக அமைகின்றது.

5.2.2.2 கிரியாவின் தற்காலத்தமிழ் அகராதி

கிரியாவின் தற்காலத்தமிழ்ச் சொற்களஞ்சியம் மொழியியல் வல்லுநர்களால் மொழியியல் கோட்பாட்டைப் பயன்படுத்தி உருவாக்கபட்டது. இவ்வகராதியில் தரப்பட்ட விளக்கங்கள் மிகத்தெளிவாகவும் மொழியியல் மற்றும் அகராதியியல் கோட்பாடுகளின் அடிப்படையில் அமைகின்றது. எனவே இவ்வகராதி சொற்பொருண்மை மயக்க நீக்கத்திற்கு மிகவும் பயனுள்ள கருவியாக அமையும்.

5.2.3 கருத்துரு வரைபட அகராதியைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்தல்

இவ்வாய்வேட்டில் மொழியியல் தகவல்களைப் பயன்படுத்தி சொற்பொருண்மை மயக்கநீக்க செய்வதற்கான முயற்சி எடுக்கப்பட்டுள்ளது. இதற்கென்று கருத்துரு வரைபட அகராதி மாதிரியாக உருவாக்கப்பட்டுள்ளது. சொற்கள் அவை உணர்த்தும் பொருண்மை அடிப்படையில் கருத்துருக்களாகப் பிரிக்கப்படும். ஒரு கருத்துரு வரைபடத்தில் அக்கருத்துரு குறித்த எல்லா தகல்களும் வெளிப்படுத்தப்பட்டிருக்கும். இதற்கு புஸ்தெஜொவ்ஸ்கியின் ஆக்கமுறை அகராதி கோட்பாட்டிலிருந்து சில செய்திகள் இங்கு பயன்படுத்தப்பட்டுள்ளன. கருத்துருவரைபட அகராதி கருத்துருக்களுக்கான விளக்கம் மட்டுமன்றி அக்கருத்துகுறித்த எல்லாச் செய்திகளையும் தரும். ஒரு விதத்தில் கூறினால் இது ஒரு களஞ்சியச் செய்திகளை உள்ளடக்கும்.

புஸ்தொஸ்கி கோட்பாடு அடிபடையில் ஒரு சொலுக்குச் பின்வரும் சொற்பொருண்மை தகவல்கள் தரப்படவேண்டும்.

- உறுப்பு உறவு (constitutive relation): ஒரு பொருளுக்கும் (object) அதன் உறுப்புப் பாகங்களுக்கும் (constituent parts) இடையில் உள்ள உறவு.
- வடிவ/முறை உறவு (formal relation): பொருளை (object) பெரிய சொற்களத்திற்குள் வேறுபடுத்தம் செய்கிறது.
- நோக்க உறவு (telic relation): பொருளின் தேவையையும் செயல்பாட்டையும் தருகிறது.
- செயலி உறவு (agentive relation): பொருளின் மூலம் அல்லது அதைக் கொண்டு வருகையில் உட்படும் காரணிகளைத் தருகின்றது.

எடுதுக்காட்டாகப் புதினம் என்ற கருத்துருவை எடுத்துக்கொள்வோம். அதன் சொற்பொருண்மைத் தகவல்கள் பின்வருவனவற்றை உள்ளடக்கும்:

புதினம்	உறுப்புப்பங்களிப்பு=கதை	.....
	வடிவப் பங்களிப்பு=புத்தகம்	
குண அமைப்பு =	நோக்கப் பங்களிப்பு=படிப்பது	
	செயலிப் பங்களிப்பு=எழுவது	

புத்தகம் என்ற கருத்துருவை எடுத்துக்கொண்டால் அது குறைந்தது பின்வரும் செய்திகளால் விளக்கப்படவேண்டும்.

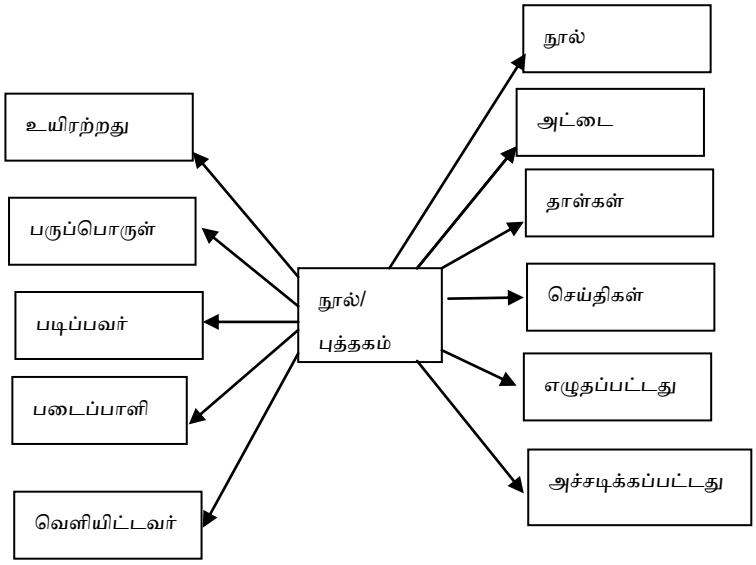
அதன் உறுப்புப் பங்களிப்பு = அது படிப்பதற்கான ஒரு பொருள்

அதன் வடிவப் பங்களிப்பு = அது தாள்களால் ஆனது, அட்டை போடப்பட்டது, படிப்பதற்கான செய்திகள் அடங்கியது

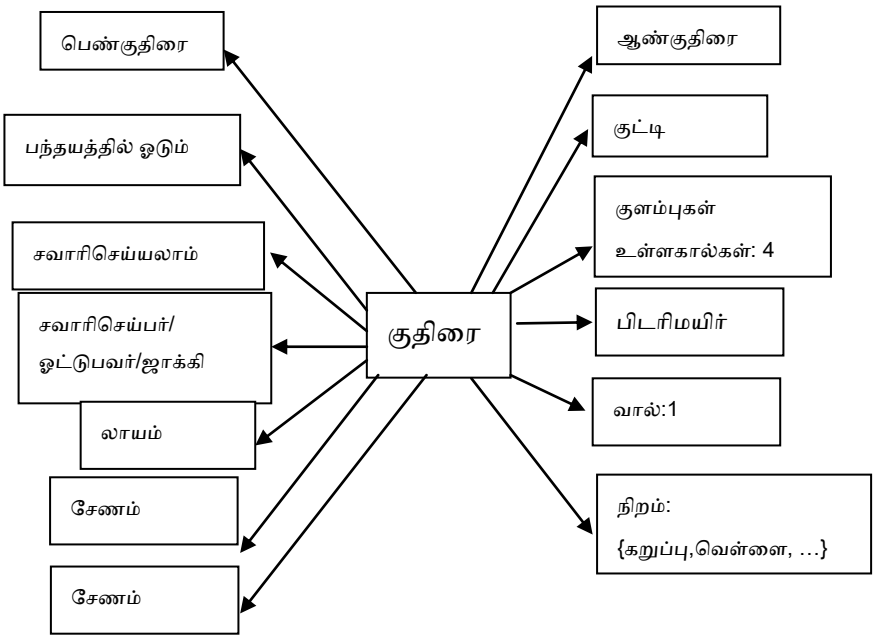
நோக்கப் பங்களிப்பு = படிப்பது

செயலி பங்களிப்பு = எழுதுவது

இதன் படி *புத்தகம்* என்பது பின்வரும் விளக்கத்தைப் பெறும்: படிப்பதற்குப் பயன்படும் எழுதப்பட்ட, வெளியிடப்பட்ட தாளகளால் ஆன, அட்டை போடப்பட்ட, படிப்பதற்கான செய்திகள் அடங்கிய ஒரு திடப்பொருள். இதைப் பின்வருமாறு கருத்துரு வரைபடமாகத் தரல்லாம். கருத்துருக்களுக்கு இடையே உள்ள உறவுகள்தரப்படவில்லை.



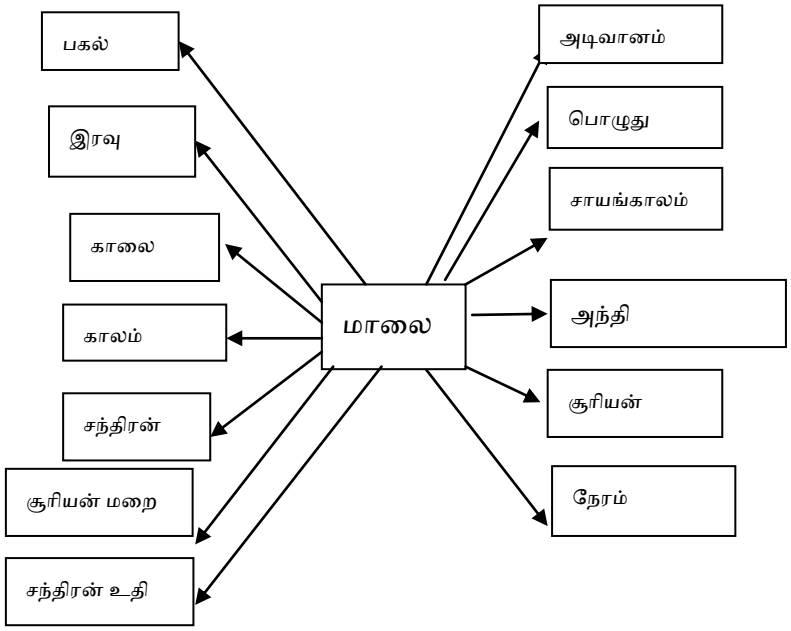
இதுபோல் *குதிரை* என்ற கருத்துருவை எடுத்துக்கொண்டால் அது பின்வருமாறு தகவல்களை வேண்டும்



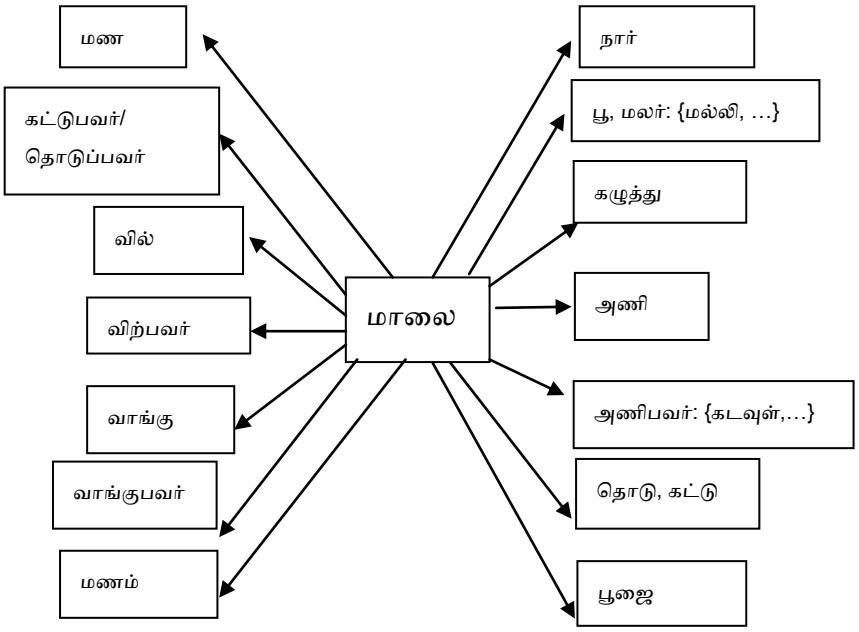
இங்கு முன்மொழிகின்ற கருத்துரு வரைபட அகராதி குறைந்தது மேற்கண்ட தகவல்களைக் கொண்டிருக்க வேண்டும்.

இக்கருத்துரு வரைபட அகராதியைக் கொண்டு பாஸ்கரன் அவர்கள் சுட்டிக்காட்டிய சொற்பொருள் மயக்கம் உள்ள சொற்களின் பொருண்மை மயக்கத்தையும் பிற சொற்களின் பொருள்மயக்கங்களையும் நீக்கலாம். எடுத்துக்காட்டாக *மாலை1* (பொழுது), (கழுத்தில் போடும்) *மாலை2* என்ற கருத்துருவின் கருத்துரு வரைபட அகராதி ஒரளவுக்குப் பின்வருமாறு அமையும்.

மாலை1


பகல்
இரவு
காலை
காலம்
சந்திரன்
சூரியன் மறை
சந்திரன் உதி
மாலை
அடிவானம்
பொழுது
சாயங்காலம்
அந்தி
சூரியன்
நேரம்

மாலை2



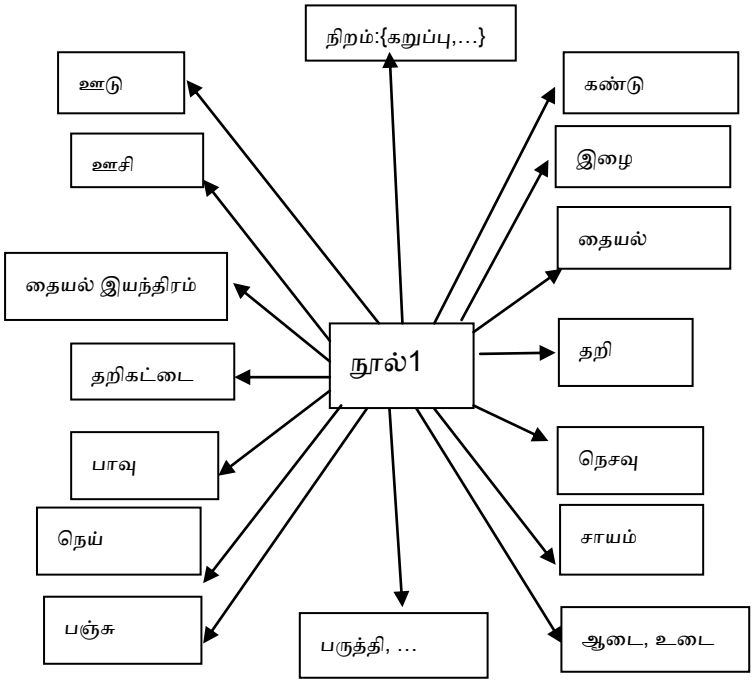
மேற்கண்ட *மாலை1*, *மாலை2* என்ற கருத்துரு வரைபட அகராதி விளக்கம் அடிப்படையில் பாஸ்கரன் அவர்களின் தரவை பின்வருமாறு பொருண்மை மயக்க நீக்கம் செயலாம்.

எண்	**வாக்கியம்**	**கருத்துரு**
1	தெரசாவின் அறப்பணிகளுக்கு நிகரில்லாத புகழ் மாலைகளாக அமைந்து மகிழ்வித்தன.	மாலை2
2	…தினந்தோறும் காலை மாலை ஆகிய இரு வேளைகளிலும்	மாலை1
3	காலையில் சூரியன் தோன்றிய நேரத்தில் இருந்து மாலையில் மறையும் நேரம் வரை	மாலை1
4	சூரியகாந்திப்பூ காலை, மாலை வேளைகளில் திசை	மாலை1
5	சுற்றுலாப் பேருந்து அடுத்த வெள்ளிக்கிழமை மாலை சென்னைக்குத்	மாலை1

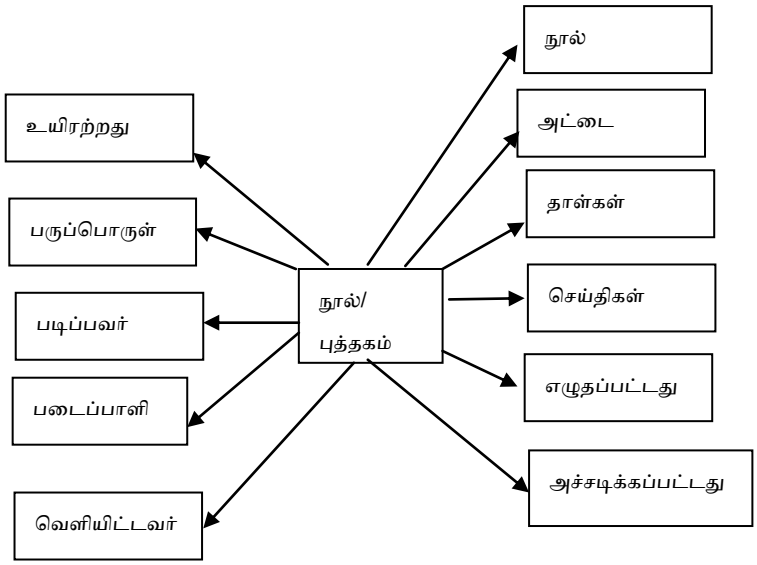
	திரும்புகிறது	
6	காலையில் 5 மணிக்கும், பிறகு 9:30 மணிக்கும், மலையில் 6:30 மணிக்கும் S1	மாலை1
7	அரசு தூக்கிலிட முடிவு செய்த நாளின் முதல் நாள் மலையே, தூக்கிலிட்டுக் S1	மாலை1
8	தத்தும் வீசிய துண்டறிக்கைகள் அன்று மாலையே "இந்துஸ்தான் டைம்ஸ்" S1	மாலை1
9	நடவடிக்கைகளின் காரணமாக நேற்று மாலையிலேயே வந்த செய்தியின்படி S1	மாலை1
10	"குரங்கு கையிலே பூமாலை" என்று அதைப்போல ஊராட்சி S2	மாலை2
11	சிலர், 'நான் மாலையையே மலர் ஆக்குவேன்	மாலை2
12	அழகு, இளமை, கருவிழி, செல்லம், வாகனம், மாலை, கொடி முதலியவை	மாலை2
13	அல்லது திருநீறு காணப்படும் சபையோர் போடும் மாலைகளைப் கழற்றாமலே	மாலை2
14	நானும் அந்தச் செய்தியை நேற்று மாலை அமைச்சரவைக் கூட்டம்	மாலை1
15	உணவுப் பொருள்களை உட்கொள்ளும் பொழுதும், மாலையில் மேய்ச்சலுக்குப் பிறகும்	மாலை1

பாஸ்கரன் அவர்கள் பொருண்மை மயக்கம் உள்ள *நூல்1* (இழை), *நூல்2* (புத்தகம்) என்ற கருத்துருக்களைக் கருத்துரு வரைபட அகராதி அடிப்படையில் பொருண்மை மயக்கம் நீக்குவதைக் கருதுவோம். *நூல்2இன்* கருத்துரு வரைபட அகராதி முன்னர் தரப்பட்டுள்ளது; இது மீண்டும் இங்கு ஒப்புமைக்காகக் காட்டப்பட்படுள்ளது. *நூல்1*-இன் கருத்துரு வரைபட அகராதி பின்வருமாறு அமையும்.

நூல்1


நிறம்:{கறுப்பு,...}
ஊடு
கண்டு
ஊசி
இழை
தையல்
தையல் இயந்திரம்
நூல்1
தறி
தறிகட்டை
நெசவு
பாவு
சாயம்
நெய்
பஞ்சு
பருத்தி, ...
ஆடை, உடை

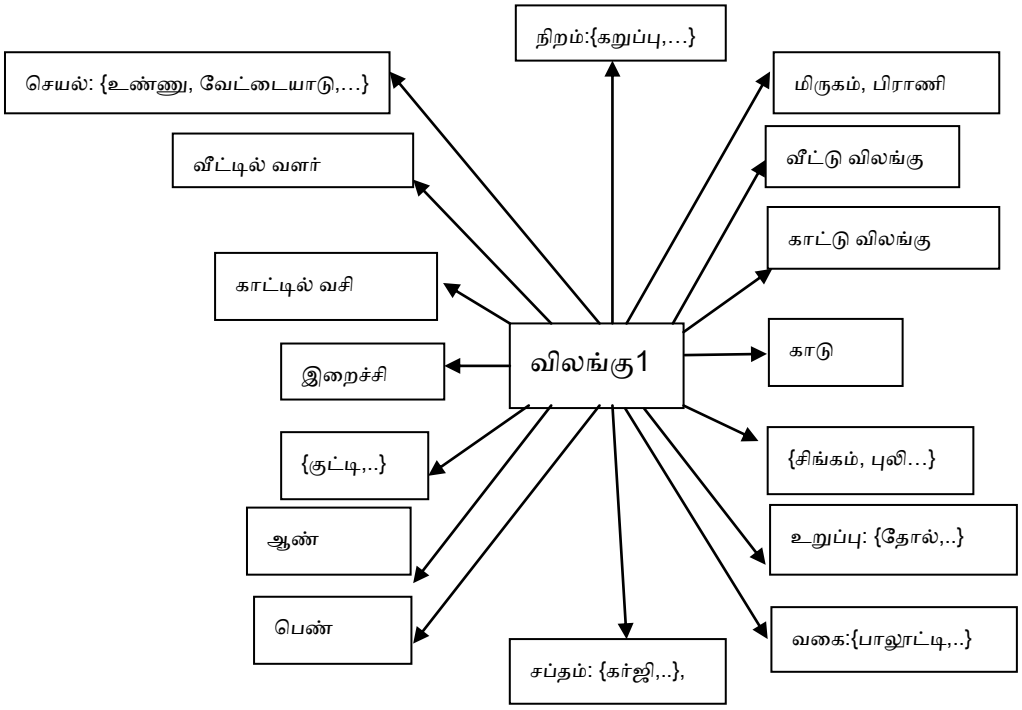
நூல்2


நூல்
உயிரற்றது
அட்டை
தாள்கள்
பருப்பொருள்
நூல்/
புத்தகம்
செய்திகள்
படிப்பவர்
எழுதப்பட்டது
படைப்பாளி
அச்சடிக்கப்பட்டது
வெளியிட்டவர்

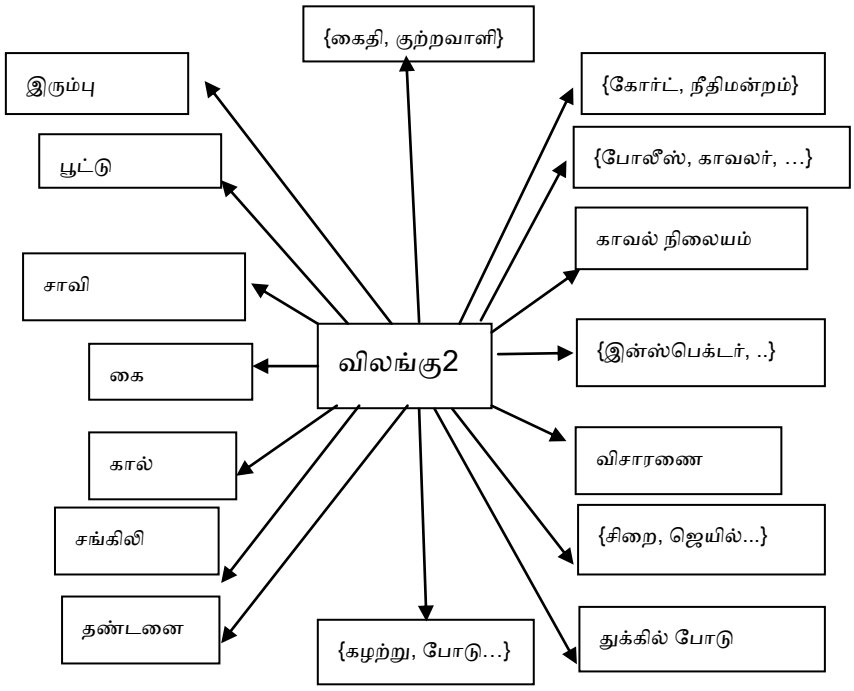
எண்	வாக்கியம்	கருத்துரு
1	… சட்டையின் நூல் இழை….	நூல்1
2	தறியில் நூல் சிக்கி துணி நெய்ய..	நூல்1
3	…நூலில் சாயம் ஏற்றி…	நூல்1
4	கதர் துணி நெய்ய நூல் தேவை…	நூல்1
5	தறியில் ஊடு…நூல்	நூல்1
6	ஊசியில் நூல் கோர்க்க இயல…	நூல்1
7	நூல் ஆசிரியர் பெயர் தெரிவில்லை	நூல்2
8	பழனியப்பா நூல் பதிப்பகத்தில் அச்சடிக்கபட்ட…	நூல்2
9	நூல் 1000 பக்கங்களைக் கொண்டு ….	நூல்2
10	நூல் 1/8 டெம்மி அளவு தாள் கொண்டு …	நூல்2

அடுத்ததாகப் பாஸ்கரன் அவர்கள் பொருண்மை மயக்கம் உள்ளவை என்று எடுத்துக்கொண்ட *விலங்கு* (மிருகம்), *விலங்கு* (கைவிலங்கு) என்ற கருத்துருக்களைக் கருத்துரு வரைபட அகராதி அடிப்படையில் பொருண்மை மயக்கம் நீக்குவதைக் கருதுவோம். *விலங்கு1*, *விலங்கு2* இவற்றின் கருத்துரு வரைபட அகராதி பின்வருமாறு அமையும்

விலங்கு1


நிறம்:{கறுப்பு,…}
செயல்: {உண்ணு, வேட்டையாடு,…}
மிருகம், பிராணி
வீட்டில் வளர்
வீட்டு விலங்கு
காட்டு விலங்கு
காட்டில் வசி
விலங்கு1
காடு
இறைச்சி
{குட்டி,..}
{சிங்கம், புலி…}
ஆண்
உறுப்பு: {தோல்,..}
பெண்
வகை:{பாலூட்டி,..}
சப்தம்: {கர்ஜி,..},

விலங்கு2



எண்	தரவு வாக்கியங்கள்	கருத்துரு
1	விலங்குகள் காட்டில் அலைந்து திரிந்து…	விலங்கு1
2	காட்டு விலங்கான சிங்கம்…	விலங்கு1
3	காட்டில் சிங்கம், புலி, யானை,…ஆகிய விலங்குகள்..	விலங்கு1
4	விலங்குகளில் பெரிதான யானையின் தந்தங்கள் விலைமிக்கவை என…	விலங்கு1
5	நாய், பூனை, பசு, குதிரை போன்ற வீட்டு விலங்குகளை…	விலங்கு1
6	போலீஸ் குற்றவாளிகளைக் கைது செய்ய…	விலங்கு2
7	கைதியைக் கோர்ட்டில் கொண்டுபோக…	விலங்கு2
8	கைதிகளை விசாரணை செய்து தண்டனை	விலங்கு2

	அளிக்க...	

உள்ளடங்கு-உள்ளடக்கு மொழி உறவுகள், எதிர்மொழிகள் உறவுகள், முழு-சினை மொழி உறவுகள், வகை உறவுகள், உறுப்பினர் உறவுகள், சொந்த பந்த உறவுகள், சமூக உறவுகள் போன்ற எல்லா உறவுகளையும் கருத்துரு வரைபட அகராதியாகத் தர இயலும்.

கருத்துரு வரைபட நெறிமுறையைப் பயன்படுத்தி சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறையின் மாதிரி உருவாக்கப்பட்டு இவ்வாய்வின் விளைவாய் தரப்பட்டுள்ளது.

5.2.4 முடிவுரை

சொற்பொருண்மை மயக்கநீக்கத்திற்குப் பல வழிமுறைகள் உள்ளன. கண்காணிக்கபட்ட அணுகுமுறைகளைக் காட்டிலும் கண்காணிக்கப்படாத அணுகுமுறைகள் சிக்கனமானது. அறிவு அடிப்படையிலான அணுகுமுறைகள் வரவேற்கப்படவேண்டியவை ஆகும்; கண்காணிக்கப்பட்ட அணுகுமுறைகளைப் போல இவ்வணுகுமுறைகளும் மொழியியல் தகவல்களைப் பயன்படுத்துகின்றன; குறிப்பாகச் சொற்பொருண்மையியல் தகவல்களைப் பயன்படுத்துகின்றன. மொழியியல் அடிப்படையில், குறிப்பாகச் சொற்பொருண்மையியல் அடிப்படையில் பொருண்மைமயக்கச் சிக்கலுக்குத் தீர்வுகாண முயல்வேண்டும். பொருண்மைமயக்கச் சிக்கல் சொற்பொருண்மை மொழியியல் அடிட்படையில் அணுகப்பட்டால் தான் தீர்வுகாண்பதற்கான வழிமுறைகளை நிறைவாக/தீர்மானமாகக் காண இயலும். பாஸ்கர் அவர்களின் கொத்து/திரட்டு நெறிமுறை கண்காணிக்கப்படாத அணுகு முறைகளாகும். இம்முறை ஓரளவுக்குப் பலன் விளைவிப்பதாகும். அவர்தம் வேற்றுமைக் குறியீட்டு முறை அவ்வளவு நம்பத்தகுந்தது அல்ல. தற்போது சொல்வலைகளைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்யும் முயற்சிகள் நடந்து வருகின்றன. இவ்வாய்வேட்டில் முன்மொழியப்பட்ட கருத்துரு வரைபடம் அடிப்படையிலான அணுகுமுறை மொழியில் மற்றும் சொற்பொருண்மையியல் கோட்பாடுகளின் அடிப்படையில் அமைகிறது. மேற்சொன்ன கருத்துரு வரைபட அகராதியை உருவாக்குவது சவாலுக்கு உரியதாகும். இம்முறை சிக்கமானதல்ல. இருப்பினும் இந்நெறிமுறை வரவேற்கத்தகுந்தாகும். கண்காணிக்கப்படாத அணுகுமுறைகளை அல்லது புள்ளியியல்சார் அணுகுமுறைகளைப் பின்பற்றும் முன்னர் அறிவு அடிப்படையிலான சொற்பொருண்மையியல் சார் அணுகுமுறைகளுக்கான ஆயத்தங்கள் மேற்கொள்ளப்படவேண்டும். ஒரு கலப்படமான,

அதாவது கண்காணிக்கப்பட்ட, கண்காணிக்கப்படாத மற்றும் அறிவு அடிப்படையிலான அணுகுமுறைகளை இணைத்துச் செய்யப்படும் அணுகுமுறை நம்பத்தகுந்ததாயும் தீர்வுகாண்பதாயும் சிறப்பாகவும் அமைய இயலும்.

இயல் 6

முடிவுரை

இவ்வியலில் 2, 3, 4, 5 ஆகிய இயல்களில் கூறப்பட்டுள்ள செய்திகள், தகவல்கள், வாதங்கள் என்பனவற்றின் சுருக்கமும் இவ்வாய்வால் அறியப்பட்ட உண்மைகளும் கண்டுபிடிப்புகளும் பயன்பாடுகளும் இவ்வாய்வின் எதிர்கால நடவடிக்கைகளும் கூறப்பட்டுள்ளன.

6.1 ஆய்வுச்சுருக்கம்

பொருண்மை மயக்கதை அமைப்புப் பொருண்மை மயக்கம், சொற்பொருண்மை மயக்கம் என இரண்டாகப் பகுக்க இயலும். சூழல் இன்றி பேசப்படும் எந்த வாக்கியமுமே பொருண்மை மயக்கம் காட்டும். ஒரு வாக்கியம் அதன் அமைப்பு காரணமாகவும் சொற்பொருண்மை காரணமாகவும் பொருண்மை மயக்கத்தை வெளிப்படுத்தும். பேசுச் சுழலிலும் எழுத்துச் சூழலிலும் இது நிகழலாம். எந்த வாக்கியத்தையும் சூழல் இல்லாமல் சரியாகப் பொருள்கொள்ள இயலாது. மொழிச் சூழல் (language context) (தொடரியல் சூழல், சொற்சூழல்), கருத்தாடல் சுழல் (discourse context) வாக்கியப் பொருண்கோண்மையில் பெரும் பங்கு வகிக்கின்றன. கருத்தாடல் சூழல் மட்டுமன்றி பயன்வழியியல் சூழலும் (pragmatic context) வாக்கியப் பொருள்கோளில் பங்களிப்பு செய்கின்றது. சுழலில்லா வாக்கியம் பொருள்தராது எனலாம். சொற்களுக்கும் இது பொருந்தும். சூழல் இல்லாமல் சொற்களுக்குப் பொருளே இல்லை எனலாம். சொற்கள் வாக்கியத்தில் பயன்படுத்தப்படும் போதுதான் அவை பொருள் பெறுகின்றன என்றுகூடக் கூறலாம். பொருண்மை மயக்கங்கள் தாய் பொழி பேசுபவர்களுக்கு அதிகச் சிக்கலைத் தராது. ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு மொழி மாற்றம் செய்யும் போது இத்தகைய மயக்கங்கள் சிக்கல்களை விளைவித்தன. மனித மொழிபெயர்ப்பாளர்கள் மொழிச் சூழல், கருத்தாடல் சூழல், பயன்வழியியல் சூழல் ஆகிய மூன்று சூழல்களையும் அறிந்து கொண்டு அவற்றின் அடிப்படையில் சொற்களுக்கும் வாக்கியத்திற்கும் பொருள்கோள் செய்கின்றனர். ஆனால் கணிப்பொறிக்கு இச்சூழல்களைப் பற்றிய அறிவு இல்லாததன் காரணமாக அவை பொருண்மை மயக்கச் சிக்கல்களுக்கு ஆளாகின்றன. கணிப்பொறிக்கு மனித அறிவைச் செயற்கையாகத் தந்தால் தான் அவை பொருண்மை மயக்கங்களைப் புரிந்துகொண்டு மொழிபெயர்ப்பில் ஈடுபட இயலும். இதற்கான முயற்சி பல காலகட்டங்களில் மேற்கொள்ளப்பட்டு நடைமுறைப் படுத்தப்பட்டு

வருகின்றன. பொருண்மை மயக்கச் சிக்கல்களின் பரிமாணம் காரணமாக இன்றும் இத்தகைய முயற்சிகள் நடைபெற்று வருகின்றன.

சொற்பொருண்மை மயக்கம் சொல்வகைப்பாட்டுப் பொருண்மை மயக்கம், ஒப்புருச்சொன்மைசார் பொருண்மை மயக்கம், பல்பொருள் ஒருமொழியம்சார் பொருண்மை மயக்கம், மாற்றப் பொருண்மை மயக்கம் என வகைப்படுத்தப்படும். அமைப்புப் பொருண்மை மயக்கம் உண்மையான அமைப்புப் பொருண்மை மயக்கம், யதேட்சையான அமைப்புப் பொருண்மை மயக்கம் என வகைப்படுத்தப்பட்டும். அமைப்புப் பொருண்மை மயக்கத்தை நீக்குவதற்கான வழிமுறைகள் ஆய்வுக்குரியன; இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் பல நடைமுறைகளைக் கையாளுகின்றன. பிற பொருண்மை மயக்கங்களான குறிப்பொருள் பொருண்மை மயக்கம், நோக்கப் பொருண்மை மயக்கம் போன்றனவையும் கருத்தில் கொள்ளப்படவேண்டும். சுருக்கமாகக் கூறினால் சொற்பொருண்மை மயக்கநீக்கம் மிகக்கடினமான சவாலாகும். இதை நேரிடுவது என்பது போரிடுவதற்குச் சமமாகும்.

மொழியியலானது கள ஆய்வில் கிடைக்கும் தரவுகளைத் தான் மொழி ஆய்வுக்குப் பயன்படுத்தி வந்தது. சாம்ஸ்கி தரவு அடிப்படையிலான ஆய்வு அல்லது உற்றுநோக்கல் அடிப்படையான ஆய்வின் முடிவுகள் குறையுள்ளவை எனக்கூறி உள்ளுணர்வு அடிப்படையிலான ஆய்வை ஊக்கப்படுத்தினார். மொழியை விதிகளாகவும் மொழி இலக்கணத்தை அறிதிறனை வெளிப்படுத்தும் அக அமைப்பாகவும் கருத அவர் கருத்துரைத்தார். கணினியின் வரவால் மொழியியல் ஆய்வுகள் கணிப்பொறி தழுவியதாய் மாறியபோது மொழியியல் ஆய்வுகள் பண்பு மொழியியலிலிருந்து (qualitative linguistics) விலகி அளவுசார் மொழியியலுக்குத் (statistical linguistics) தாவியது. இத்தகைய தாவலைப் பொருண்மை மயக்க ஆய்வுகளிலும் காணலாம். அமைப்பு மொழியியலில் பொருண்மை மயக்கங்கள் அமைப்புப் பொருண்மை மயக்கமாகவும் சொற்பொருண்மை மயக்கமாகவும் பகுக்கப்பட்டு விளக்கங்கள் தரப்பட்டன. சாம்ஸ்கியின் மாற்றிலக்கணக் கோட்பாடுகள் மயக்கமுள்ள வாக்கியங்களுக்கு வேறுபட்ட அக அமைப்புகளைத் தந்து ஒற்றுமையான புற அமைப்புகளைப் பெறுவதற்கான மாற்றுவிதிகள் பற்றி விவரித்தது. இத்தகைய விளக்கங்கள் மனித் மூளைக்கு ஏற்றவாறு அமைந்தனவே ஒழிய கணிப்பொறிகொண்டு இச்செயல்பாடுகளைப் புரியச்செய்யவைத்து பொருண்மை மயக்கநீக்கம் செய்வதற்கான வழிமுறைகள் இயலாது போயிற்று. கணிப்பொறியைப் பொறுத்தமட்டில்

பொருண்மை மயக்கநீக்க அறிவு கணிப்பொறிக்குக் கற்றல் பயிற்சிகளாகத் தரப்பட்டு புரியவைக்கப்பட்டன. இந்தகைய கற்றல் பயிற்சிகளின் பின்னணியிலான சொற்பொருண்மை வழிமுறைவரைவுகள் பின்வருமாறு அமைந்தன:

1. கண்காணிக்கப்பட்ட இயந்திரம் கற்றல் (supervised learning)
2. பகுதி கண்காணிக்கப்பட்ட இயந்திரம் கற்றல் (semi-supervised learning)
3. கண்காணிக்கப்படாத இயந்திரம் கற்றல் (unsupervised learning)
4. அறிவு அடிப்படையிலான இயந்திரம் கற்றல் (knowledge based learning)

தானியக்க மொழிப் பகுப்பாய்வு (automatic language processing) இருக்கிறது வரை சொற்பொருண்மை மயக்கநீக்க ஆய்வுகள் ஒரு வரலாறாக வந்து கொண்டிருக்கும். பின்னோக்கிப் பார்த்தால் பெரும்பாலான சிக்கல்களும் அவற்றிற்கான அணுகுமுறைகளும் தொடக்ககாலத்திலேயே புரிந்துகொள்ளப்பட்டுள்ளன என்பது குறிப்பிடத்தக்கதாகும். சொற்பொருண்மை மயக்கநீக்கத்தின் மீதான தொடக்ககால ஆய்வுகள் ஒத்தறி அடிபடையில் மறைக்கப்பட்ட/மறக்கப்பட்ட நூல்களாகவும் கட்டுரைகளாகவும் பல ஆய்வுக்களங்களிலும் துறைகளிலும் வெளிவந்ததன் காரணமாகத் தற்போதைய ஆய்வாளர்களும் படைப்பாளிகளும் அவற்றைப் பற்றி அறியாதிருக்கின்றனர் என்றுகூற இயலும். கடந்த 50 வருடங்களில் சொற்பொருண்மை மயக்கநீக்க ஆய்வில் மிகக்குறைவான முன்னேற்றமே எய்தப்பட்டுள்ளது என்பது ஆச்சரியப்படத்தக்கதாகும். அண்மைக்கால ஆய்வுப்பணிகள் 90% அல்லது அதற்கு மேற்பட்ட விளைவுகளை/முடிவுகளைக் குறிப்பிட்டாலும் இவ்வாய்வுகள் எடுத்துக்காட்டாகச் சில சொற்களையே, பெரும்பாலும் பெயர்ச் சொற்களையே உட்படுத்தியுள்ளன அல்லது எடுத்தாண்டுள்ளன; அதிலும் பரந்த அர்த்த வேறுபாடுகளையே செய்துள்ளன.

சொற்பொருள் மயக்கநீக்க ஆய்வுகள் ஒரு முழு சுற்று சுற்றி தொடக்ககாலத்தில் சொற்பொருண்மை மயக்கநீக்கச் சிக்கல்களுக்குத் தீர்வுகாணப் பயன்படுத்தப்பட்ட அனுபவாத அணுகு முறைகளுக்கும் தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகளுக்கும் மீண்டும் வந்துள்ளன. 1990களில் போதுமான அளவிலான மூலவளங்கள் மற்றும் மேம்படுத்தப்பட்ட புள்ளியியல் அணுகுமுறைகள் காரணமாகத் தொடக்ககால முடிவுகளிலும் விளைவுகளிலும் முன்னேற்றம் ஏற்பட்டுள்ளன; இருப்பினும் நாம் தற்போதைய சட்டகத்தில் எவ்வளவு சாதிக்க இயலுமோ அதன்

எல்லையை அடைந்துவிட்டோம். இதன் காரணமாக நாம் சொற்பொருண்மை மயக்கநீக்க ஆய்வுகளின் தற்போதைய நிலையைப் பற்றி மதிப்பிடுவது அவசியமாகும்.

சொற்பொருண்மை மயக்கநீக்கத்தின் கடினம் எடுத்துக்கொள்ளப்பட்ட அர்த்தங்களின் நுணுக்கத்தைச் சார்ந்திருக்கின்றது. யரோவ்ஸ்கி மற்றும் பிறர் (Yarowsky [1995] and Stevenson and Wilks [2001]) 95%க்கு அண்மையான அல்லது அதற்கு அதிகமான துல்லியம் ஒருசொல்போலிகளின் பொருண்மை மயக்கநீக்கத்தில் பெற இயலும் என்று காட்டியுள்ளனர். ஆனால் பல்பொருண்மையின் பொதுவான கருத்துச்சாயலுக்கு வருகையில் சிக்கல் கடினமாகின்றது; இங்கு நுண்மையானது, பொண்மை மயக்கநீக்க ஒழுங்குமுறைகளின் செயல்பாட்டிலும் மனித அடையாளப்படுத்துபவர்களுக்கு இடையில் உடன்பாட்டிலும் வேறுபாட்டை உருவாக்குகின்றது.

கண்காணிக்கப்பட்ட நெறிமுறைகள் பிற நெறிமுறைகளைக் காட்டிலும் எந்தவித ஐயமுமின்றி நன்றாகச் செயலாற்றுகின்றது. இருப்பினும் வேறுபட்ட பொருட்புலங்களுக்கு, மொழிகளுக்கு, பணிகளுக்கு அதிக அளவிலான பயிற்சித் தரவுத்தொகுதிகள் கிடைப்பதைச் சார்ந்திருப்பது என்பது சாத்தியமான அனுமானம் அல்ல. Ng [1997] மிக உயர்ந்த துல்லியமான பரந்த செயலெலையைக் கொண்ட பொருண்மைமயக்கநீக்க ஒழுங்குமுறையைப் பெற/உருவாக்க 3.2 மில்லியன் அர்த்தம் அடையாளப்படுத்தப்பட்ட சொற்கள் தேவை என்று நிர்ணையித்துள்ளார்.

மாறாக அறிவு அடிப்படையிலான அணுகுமுறைகள் பலகாரணங்களால் சிறிய-மத்திய செயல்பாடுகளுக்கு மிகவும் நம்பகமானவை: முதலாவது கூடுதல் அறிவு இருந்தால்/கிடைத்தால் கூடுதல் செயல்பாடு கிடைக்கும். [Cuadros and Rigau 2006; Navigli and Lapata 2007]; இரண்டாவது அவர்கள் சார்ந்திருக்கும் மூலவளங்கள் ஏறுமுகமாகச் செறிவூட்டலைப் பெறுகின்றன (எடுத்துக்காட்டாக சொல்வலைகள் சொல்வலைபிளஸ் என்பன); மூன்றாவது பொருண்மை வலையின் பயன்பாடுகள் ஆற்றல் மிக்கப் பொருட்புல மூலப்பொருண்மையியல் ஆய்வுகளின் ஆற்றலைப் பயன்படுத்தும் மற்றும் பயன்படுத்துபவர்கள், தொழில்கள், மற்றும் ஒழுங்குமுறைகள் இவற்றிற்கிடையில் பொருண்மை ஊட்டாட்டத் திறனை அறிவுச் செழுமிய நெறிமுறைகள் வேண்டுகின்றன.

6.2 ஆய்விலிருந்து அறியப்பட்ட உண்மைகளும் கண்டுபிடிப்புகளும்

சொற்பொருண்மை மயக்கநீக்கத்தின் ஆய்வு 1950களிலிருந்து நடந்துகொண்டிருக்கின்றது. சொற்பொண்மை மயக்கநீக்கம் கடினமான ஆய்வுப்பணியாகும். இது மொழியின் முழுமையான கலவைத் தன்மைகளைக் கையாளுகின்றது. அமைப்பாக்கம் செய்யப்படாத மூலப் பனுவல்களிலிருந்து பொருண்மை அமைப்பை அடையாளம் காண்பதைக் குறிக்கோளாகக் கொண்டுள்ளது.

சொற்பொருண்மை மயக்கநீக்கத்திற்குப் பல வழிமுறைகள் உள்ளன. கண்காணிக்கபட்ட அணுகுமுறைகளைக் காட்டிலும் கண்காணிக்கப்படாத அணுகுமுறைகள் சிக்கனமானது. அறிவு அடிப்படையிலான அணுகுமுறைகள் வரவேற்கப்படவேண்டியவை ஆகும்; கண்காணிக்கப்பட்ட அணுகுமுறைகளைப் போல இவ்வணுகுமுறைகளும் மொழியியல் தகவல்களைப் பயன்படுத்துகின்றன; குறிப்பாகச் சொற்பொருண்மையியல் தகவல்களைப் பயன்படுத்துகின்றன. மொழியியல் அடிப்படையில், குறிப்பாகச் சொற்பொருண்மையியல் அடிப்படையில் பொருண்மைமயக்கச் சிக்கலுக்குத் தீர்வுகாண முயல்வேண்டும். பொருண்மைமயக்கச் சிக்கல் சொற்பொருண்மை மொழியியல் அடிட்படையில் அணுகப்பட்டால் தான் தீர்வுகாண்பதற்கான வழிமுறைகளை நிறைவாக/தீர்மானமாகக் காண இயலும். பாஸ்கர் அவர்களின் கொத்து/திரட்டு நெறிமுறை கண்காணிக்கப்படாத அணுகு முறைகளாகும். இம்முறை ஓரளவுக்குப் பலன் விளைவிப்பதாகும். அவர்தம் வேற்றுமைக் குறியீட்டு முறை அவ்வளவு நம்பத்தகுந்தது அல்ல. தற்போது சொல்வலைகளைப் பயன்படுத்தி பொருண்மை மயக்கநீக்கம் செய்யும் முயற்சிகள் நடந்து வருகின்றன. இவ்வாய்வேட்டில் முன்மொழியப்பட்ட கருத்துரு வரைபடம் அடிப்படையிலான அணுகுமுறை மொழியில் மற்றும் சொற்பொருண்மையியல் கோட்பாடுகளின் அடிப்படையில் அமைகிறது. மேற்சொன்ன கருத்துரு வரைபட அகராதியை உருவாக்குவது சவாலுக்கு உரியதாகும். இம்முறை சிக்கமானதல்ல. இருப்பினும் இந்நெறிமுறை வரவேற்கத்தகுந்தாகும். கண்காணிக்கப்படாத அணுகுமுறைகளை அல்லது புள்ளியியல்சார் அணுகுமுறைகளைப் பின்பற்றும் முன்னர் அறிவு அடிப்படையிலான சொற்பொருண்மையியல் சார் அணுகுமுறைகளுக்கான ஆயத்தங்கள் மேற்கொள்ளப்படவேண்டும். ஒரு கலப்படமான, அதாவது கண்காணிக்கப்பட்ட, கண்காணிக்கப்படாத மற்றும் அறிவு அடிப்படையிலான அணுகுமுறைகளை இணைத்துச் செய்யப்படும் அணுகுமுறை நம்பத்தகுந்ததாயும் தீர்வுகாண்பதாயும் சிறப்பாகவும் அமைய இயலும்.

கருத்துரு வரைபட நெறிமுறையைப் பயன்படுத்தி ஒரு மாதிரி சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறை (Conceptutal graph based WSD Model) உருவாக்கப்பட்டுள்ளது. இது கருத்துரு வரைபட அகராதியில் தரப்பட்டுள்ள சொற்களைக் கருத்துரு அடிப்படையில் சொற்பொருண்மை மயக்கநீக்கம் செய்யும்.

6.3 ஆய்வின் பயன்பாடுகள்

ஆய்வின் பயன்பாடுகளைப் பட்டியலிடுவது ஒரு நீண்ட பட்டியலில் விளையும். பொருண்மை மயக்கநீக்கம் இயற்கை மொழி ஆய்வில் பெரும் பங்கு வகிக்கின்றது. இயந்திர மொழிபெயர்பு ஒழுங்கு முறை, தகவல் மீட்பு ஒழுங்குமுறை, தகவல் தேடல் ஒழுங்குமுறை, கேள்வி-விடை ஒழுங்குமுறை போன்றவற்றில் சொற்பொருண்மை மயக்கநீக்கம் இன்றியமையாத பங்களிப்பு செய்கின்றது. இயந்திர மொழிபெயர்ப்பைப் பொறுத்தவரையில் தானியக்கமான சொற்பொருண்மை மயக்கநீக்கம் வரவேற்பதற்குரியது. ஆனால் தற்போதைய நிலையில் அது ஒரு கனவு எனலாம். சொற்பொருண்மை மயக்கநீக்க வழிமுறைகள்/நெறிமுறைகள் வழி தமிழ் பனுவல்களையும் உரைகளையும் பொருண்மைமயக்கநீக்கம் செய்வது தமிழ்சார் இயற்கை மொழி ஆய்வுகளுக்குப் பெரிதும் உதவும். குறிப்பாக இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கான இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைகளில் இதன் பயன் சொல்லற்கரியதாகும். தமிழுக்கு இத்தகைய சொற்பொருண்மை மயக்கநீக்க ஒழுங்குமுறை ஒரு வரப்பிரசாதமாக அமையும்.

6.4 எதிர்கால நடவடிக்கைகளும்

சொற்பொருள் மயக்கநீக்கத்திற்காக மேற்கொள்ள வேண்டிய எதிர்கால நடவடிக்கைகள் பலவாகும். குறிப்பாகக் கணினிசார் சொற்பொருண்மையியல் அடிப்படையில் ஆய்வுகள் மேற்கொள்ளப்படவேண்டும். தமிழுக்குக் கருத்துரு வரைபட அகராதி முழுமையாக உருவாக்கப்படவேண்டும். தமிழ்ச் சொல்வலை உருவாக்க ஆய்வுத்திட்டம் தமிழ்ப் பல்கலைக்கழகத்தில் தொடங்கப்பட்டு நடைபெற்று வருகின்றது. தமிழ்ச் சொல்வலை ஆய்வுத்திட்டம் சொற்களின் பொருண்மைகளைக் கருத்துருக்களாகப் பிரித்து மூலப்பொருண்மையியல் ஆய்வின் பின்னணியில் சொல் உறவுகள், பொருண்மை உறவுகள் ஆகியவற்றைக் கருத்தில் கொண்டு அகராதி நுட்பங்களையும் பொருட்புல அகராதி நுட்பங்களையும் ஒருங்கிணைத்து சொல்சார் தரவு மையம் (lexical database) உருவாக்கும் முயற்சியில் ஈடுபட்டுள்ளது. இம்முயற்சி முழுமை பெறும் போது சொற்பொருண்மை

மயக்கநீக்கத்திற்கான அறிவுச்செறிவு அல்லது அறிவு மூலவளம் கிடைக்கபெறும். தமிழ்ச் சொல்வலையை வைத்துகொண்டு அறிவு அடிப்படையிலான சொற்பொருண்மை மயக்கநீக்கம் செய்வதை முயற்சிக்கலாம். தமிழ்ச் சொல்வலையில் தரப்பட்டுள்ள கருத்துருக்கள் அடிப்படையிலான சொற்பொருண்மைசார் விளக்கங்களை வைத்துக்கொண்டு கருத்துரு வரைபட அகராதியை உருவாக்கி சொற்பொருண்மை மயக்கத்தை நீக்கலாம்.

துணை நூற்பட்டியல்

இராசேந்திரன், ச. 1999. பொருட்புல வகைப்பாடும் சொற்களஞ்சியமும் [Classification into semantic domains and Thesaurus]. புலமை தொகுதி 25, எண்2, டிசம்பர், 1999, 47-66.

இராசேந்திரன், ச. 2001 தற்காலத் தமிழ்சொற்களஞ்சியம். தமிப்பல்கலைக் கழகம், தஞ்சாவூர்.

இராசேந்திரன், ச. 2006. தமிழ் மின்சொற்களஞ்சியம். தமிழ்ப்பல்கலைக் கழகம், தஞ்சாவூர்.

க்ரியாவின் தற்காலத் தமிழ் அகராதி. க்ரியா, சென்னை.

இராசேந்திரன், ச.. சொற்பொருண்மை மயக்கநீக்கம்: கணினிசார்அணுகுமுறை [A computational approach to word sense disambiguation]. Uploaded in Academia.edu.

ABNEY, S. 2004. Understanding the Yarowsky algorithm. Computat. Ling. 30, 3, 365–395.

ABNEY, S. AND LIGHT, M. 1999. Hiding a semantic class hierarchy in a Markov model. In Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing (College Park, MD). 1–8.

ADRIAENS, Geert. 1986. Word expert parsing: A natural language analysis program revised and applied to Dutch. Leuvensche Bijdragen, 75(1):73-154.

ADRIAENS, G. 1987. WEP (word expert parsing) revised and applied to Dutch. In Proceedings of the 7th European Conference on Arti~'cial Intelligence, ECAI'86, pages 222-235, Brighton, United Kingdom, July. Reprinted in B. Du Boulay, D. Hogg, L. Steels, editors, Advances in Artij~'cial Intelligence II, pages 403-416, Elsevier.

ADRIAENS, G. 1989. The parallel expert parser: A meaning-oriented, lexically guided, parallel-interactive model of natural language understanding. In Proceedings of the International Workshop on Parsing Technologies, pages 309-319, Carnegie-Mellon University.

ADRIAENS, G. and SMALL. S.L. 1988. Word expert revisited in a cognitive science perspective. In Steven Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and ArtiJicial Intelligence. Morgan Kaufman, San Mateo, CA, pages 13-43.

AGIRRE, E., ANSA,O.,MARTINEZ, D., ANDHOVY, E. 2001. EnrichingWordNet concepts with topic signatures. In Proceedings of the NAACL Workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations (Pittsburg, PA). 23–28.

AGIRRE, E. AND EDMONDS, P., Eds. 2006. Word Sense Disambiguation: Algorithms and Applications. Springer, New York, NY.

AGIRRE, E. AND LOPEZ DE LACALLE, O. 2003. Clustering WordNet word senses. In Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP, Borovets, Bulgaria). 121–130.

AGIRRE, E. AND LOPEZ DE LACALLE, O. 2007. UBC-ALM: Combining k-nn with SVD for WSD. In Proceedings of the 4th InternationalWorkshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 342–345.

AGIRRE, E., MAGNINI, B., DE LACALLE, O., OTEGI, A., RIGAU, G., AND VOSSEN, P. 2007a. Semeval-2007 task 01: Evaluating WSD on cross-language information retrieval. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 1–6.

AGIRRE, E., M`ARQUEZ, L., AND WICENTOWSKI, R., Eds. 2007b. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval). Association for Computational Linguistics, Prague, Czech Republic.

AGIRRE, E. AND MARTINEZ, D. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the 18th International Conference on Computational Linguistics (COLING, Saarbr ¨ ucken, Germany). 11–19.

AGIRRE, E. ANDMARTINEZ, D. 2001. Learning class-to-class selectional preferences. In Proceedings of the 5th Conference on Computational Natural Language Learning (CoNLL, Toulouse, France). 15–22.

AGIRRE, E.,MART´ıNEZ,D., LOPEZ DE LACALLE, O., AND SOROA,A. 2006. Two graph-based algorithms for state-ofthe- art wsd. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (Sydney, Australia). 585–593.

AGIRRE, E. AND RIGAU, G. 1996. Word sense disambiguation using conceptual density. In Proceedings of the 16th International Conference on Computational Linguistics (COLING, Copenhagen, Denmark). 16–22.

AGIRRE, E. AND SOROA, A. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 7–12.

AGIRRE, E. AND STEVENSON, M. 2006. Knowledge sources for WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 217–251.

ALEMAN-MEZA, B., NAGARAJAN, M., RAMAKRISHNAN, C., DING, L., KOLARI, P., SHETH, A., ARPINAR, I., JOSHI, A., AND FININ, T. 2006. Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection. In Proceedings of the 15th International Conference on World Wide Web (WWW, Edinburgh, Scotland, U.K.). 407–416.

ALPAYDIN, E. 2004. Introduction to Machine Learning. MIT Press, Cambridge, MA.

AMSLER, R. A. 1980. The structure of the Merriam-Webster pocket dictionary. Ph.D. dissertation. University of Texas at Austin, Austin, TX.

ANTHONY, E. 1954. An exploratory inquiry into lexical clusters. American Speech, 29(3):175-180.

ANAYA-S´ANCHEZ, H., PONS-PORRATA, A., AND BERLANGA-LLAVORI, R. 2007. TKB-UO: Using sense clustering for wsd. In Proceedings of the 4th InternationalWorkshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 322–325.

ARTILES, J., GONZALO, J., AND SEKINE, S. 2007. The Semeval-2007 WEPS evaluation: Establishing a benchmark for the Web people search task. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 64–69.

ASPREJAN, J.D. 1974. Regular polysemy. Linguistics, 142:5-32.

ATKINS, S. 1993. Tools for computer-aided corpus lexicography: The Hector project. Acta Linguistica Hungarica 41, 5–72.

ARULMOZHI. P, SOBHA. L. (2006).Semantic Tagging for Language Processing. presented in 34th All India conference for Dravidian Linguistics held at International School of Dravidian Linguistics, Trivandrum.

ARUN, M.C. “Evaluation of Word Sense Disambiguation methods applied on Tamil Corpus. PhD thesis.

BANERJEE, S. AND PEDERSEN, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico). 805–810.

BAR-HILLEL, Y. 1960. The present status of automatic translation of languages. Advan. Comput. 1, 91–163.

BARWISE, J. and Perry J.R.. 1983.Situations and Attitudes. MIT Press, Cambridge, MA.

BARZILAY, R. AND ELHADAD,M. 1997. Using lexical chains for text summarization. In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization (Madrid, Spain). 10–17.

BASILI, R., ROCCA,M. D., AND PAZIENZA,M. T. 1997. Contextual word sense tuning and disambiguation. Appl. Artific. Intell. 11, 3, 235–262.

BASKARAN, S. 2002. Semantic analyser for word sense disambiguation. MS thesis. Madras Institute of Technology, Anna University, Chennai 2002.

BASKARAN SANKARAN, VAIDEHI V, 2002, "Role of Collocations and Case-Markers inWord Sense Disambiguation: A Clustering-Based Approach". In Proceedings of IEEE International Symposium on Natural Language Processing and Knowledge Engineering 2002. Vol. I. pp. 625-630. Hammamet, Tunisia.

BASKARAN S, VAIDEHI V, 2003, "Collocation based Word Sense Disambiguation using Clustering for Tamil". Communicated to International Journal of Dravidian Linguistics, Thiruvananthapuram, India.

BEL'SKAJA, I.K. 1957. Machine Translation of Languages. Research, 10(10).

BENTIVOGLI, L. AND PIANTA, E. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. J. Nat. Lang. Eng. 11, 3, 247–261.

BERENDT, B. AND NAVIGLI, R. 2006. Finding your way through blogspace: Using semantics for cross-domain blog analysis. In Proceedings of the AAAI Spring Symposium 2006 (AAAIS) on Computational Approaches to Analysing Weblogs (CAAW, Palo Alto, CA). 1–8.

BERNARD, J. R. L., Ed. 1986. Macquarie Thesaurus. Macquarie, Sydney, Australia.

BERNERS-LEE, T., HENDLER, J., AND LASSILA. O. 2001. The semantic Web. http://www.sciam.com/article. cfm?id=the-semantic-web&page=2.

BERRY-ROGGHE, Godelieve. 1973. Thecomputation of collocations and their relevance to lexical studies. In Adam J. Aitken, Richard W. Bailey, and Neil Hamilton-Smith, editors, The Computer and Literary Studies. Edinburgh University Press, Edinburgh, UK, pages 103-112.

BLACK, E. 1988. An experiment in computational discrimination of English word senses. IBM J. Res. Devel. 32, 2, 185–194.

BLEI, D. M., NG, A. Y., AND JORDAN,M. I. 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

BLOOMFIELD, L. 1933. Language. Holt,New York.

BORDAG, S. 2006. Word sense induction: Triplet-based clustering and automatic evaluation. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL, Trento, Italy). 137–144.

BOSER, B. E., GUYON, I. M., AND VAPNIK, V. N. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual Workshop on Computational Learning Theory (Pittsburgh, PA). 144–152.

BOYD-GRABER, J., FELLBAUM, C., OSHERSON, D., , AND SCHAPIRE, R. 2006. Adding dense, weighted connections to WordNet. In Proceedings of the 3rd International WordNet Conference (Jeju Island, Korea).

BRANTS, T. AND FRANZ, A. 2006. Web 1t 5-gram, ver. 1, ldc2006t13. Linguistic Data Consortium, Philadelphia, PA.

BRIN, S. AND PAGE, M. 1998. Anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th Conference on World Wide Web (Brisbane, Australia). 107–117.

BRISCOE, E.J. 1991. Lexical issues innatural language processing. In Ewan H. Klein and Frank Veltman,editors, Natural Language and Speech.Proceedings of the Symposium on Natural Language and Speech, pages 39-68, Springer-Verlag, Berlin.

BRODY, S.,NAVIGLI, R., AND LAPATA,M. 2006. Ensemble methods for unsupervised WSD. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL, Sydney, Australia). 97–104.

BROWN, P. F.,PIETRA, S. A. D.,PIETRA, V. J. D., ANDMERCER, R. L. 1991. Word-sense disambiguation using statistical methods. In Proceedings of 29th Annual Meeting of the Association for Computational Linguistics (Berkeley, CA). 264–270.

BRUCE, R. AND WIEBE, J. 1994. Word-sense disambiguation using decomposable models. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL, Las Cruces, NM). 139–145.

BRUCE, R. ANDWIEBE, J. 1999. Decomposable modeling in natural language processing. Comput. Ling. 25, 2, 195–207.

BUDANITSKY, A. AND HIRST, G. 2006. Evaluating WordNet-based measures of semantic distance. Computat. Ling. 32, 1, 13–47.

Buitelaar, P. 1997. A lexicon forunderspecified semantic tagging. In Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?", pages 25-33,Washington, DC, April.

BUITELAAR, P. 1998. Corelex: An ontology of systematic polysemous classes. In Formal Ontology in Information Systems, N. Guarino, Ed. IOS Press, Amsterdam, The Netherlands. 221–235.

BUITELAAR, P., MAGNINI, B., STRAPPARAVA, C., AND VOSSEN, P. 2006. Domain-specific WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 275–298.

BUITELAAR, P. AND SACALEANU, B. 2001. Ranking and selecting synsets by domain relevance. In Proceedings of the NAACL Workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations (Pittsburgh, PA).

BUNKE, H. AND SANFELIU, A., Eds. 1990. Syntactic and Structural Pattern Recognition: Theory and Applications. Vol. 7. World Scientific Series in Computer Science, World Scientific, Singapore.

BYRD, R. J., Nicoletta Calzolari, Martin S. Chodorov, Judith L. Klavans, Mary S. Neff, and Omneya Rizk. 1987. Tools and methods for computational linguistics. Computational Linguistics,13(3/4):219-240.

CAI, J. F., LEE, W. S., AND TEH, Y. W. 2007. NUS-ML: Improving word sense disambiguation using topic features. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 249–252.

CALZOLARI, N. 1984. Detecting patternsin a lexical data base. In Proceedings of the l0th International Conference onComputational Linguistics, COLING'84,pages 170-173, Stanford University, CA, July.

CARDIE, C. AND MOONEY, R. J. 1999. Guest editors' introduction: Machine learning and natural language. Mach. Learn. 34, 1–3, 5–9.

CARPUAT, M. AND WU, D. 2005. Word sense disambiguation vs. statistical machine translation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL, Ann Arbor, MI). 387–394.

CARPUAT, M. ANDWU, D. 2007. Improving statistical machine translation using word sense disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL, Prague, Czech Republic). 61–72.

CHAFE, W. 1970. Meaning and Structure of Language. University of Chicago Press, Chicago

CHAFE, W. 2000. Meaning in Language: An Ingroduction to Semantics and pragmatics. Oxford University Press, Oxford, New York.

CHAN, Y. S. AND NG, H. T. 2005. Scaling up word sense disambiguation via parallel texts. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI, Pittsburgh, PA). 1037–1042.

CHAN, Y. S., NG, H. T., AND CHIANG, D. 2007a. Word sense disambiguation improves statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (Prague, Czech Republic). 33–40.

CHAN, Y. S., NG, H. T., AND ZHONG, Z. 2007b. NUS-PT: Exploiting parallel texts for word sense disambiguation in the English all-words tasks. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 253–256.

CHARNIAK, E., BLAHETA, D., GE, N., HALL, K., HALE, J., AND JOHNSON, M. 2000. Bllip 1987-89 WSJ corpus release 1. Tech. rep. LDC2000T43. Linguistic Data Consortium (Philadelphia, PA).

CHEN, J.N. AND CHANG, J.S. 1998. “Topical empirical study of smoothing techniques for language modeling.” Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 24-27 June 1996. University of California, Santa Cruz, California, 310-318.

CHKLOVSKI, T. AND MIHALCEA, R. 2002. Building a sense tagged corpus with Open Mind Word Expert. In Proceedings of ACL 2002 Workshop on WSD: Recent Successes and Future Directions (Philadelphia, PA).

CHKLOVSKI, T. AND MIHALCEA, R. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP, Borovets, Bulgaria).

CHKLOVSKI, T., MIHALCEA, R., PEDERSEN, T., AND PURANDARE, A. 2004. The Senseval-3 multilingual English-Hindi lexical sample task. In Proceedings of the 3rd

InternationalWorkshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain). 5–8.

CHKLOVSKI, T. AND PANTEL, P. 2004. Verbocean: Mining the Web for fine-grained semantic verb relations. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP, Barcelona, Spain).

CHODOROW, M., BYRD, R., AND HEIDORN, G. 1985. Extracting semantic hierarchies from a large on-line dictionary. In Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (Chicago, IL). 299–304.

CHURCH, K. W. AND RAU, L. F. 1995. Commercial applications of natural language processing. Commun. ACM 38, 11, 71–79.

CIARAMITA, M. AND ALTUN, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP, Sydney, Australia). 594–602.

CIARAMITA, M. AND JOHNSON, M. 2000. Explaining away ambiguity: Learning verb selectional restrictions with Bayesian networks. In Proceedings of the 18th International Conference on Computational Linguistics (COLING, Saarbr ¨ ucken, Germany). 187–193.

CIMIANO, P. 2006. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer, New York, NY.

CLARK, S. AND WEIR, D. 2002. Class-based probability estimation using a semantic hierarchy. Computat. Ling. 28, 2, 187–206.

CLEAR, J. 1993. The British National Corpus. In The Digital Word: Text-Based Computing in the Humanities, P. Delany and G. P. Landow, Eds. MIT Press, Cambridge, MA. 163–187.

COHN,D.,ATLAS, R., AND LADNER,R. 1994. Improving generalization with active learning. Mach. Learn. 15, 2, 201–221.

COHN, T. 2003. Performance metrics for word sense disambiguation. In Proceedings of the Australasian Language Technology Workshop (Melbourne, Australia). 49–56.

COLLINS,M. 2004. Parameter estimation for statistical parsing models: Theory and practice of distribution free methods. In New Developments in Parsing Technology,H. Bunt, J. Carroll, andG. Satta, Eds. Kluwer, Dordrecht, The Netherlands, 19–55.

COLLINS, A. M. and LOFTUS, E. E.1975. A spreading activation theory of semantic processing. Psychological Review, 82(6):407-428.

COST, S. AND SALZBERG, S. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. Mach. Learn. 10, 1, 57–78.

COTTRELL, G.W. and Small S.L.1983. A connectionist scheme for modelling word sense disambiguation.Cognition and Brain Theory, 6:89-120.

COTTRELL, G.W. 1985. A ConnectionistApproach to Word-Sense Disambiguation. Ph.D. thesis. Department of ComputerScience, University of Rochester.

COTTRELL, G. W. 1989. A Connectionist Approach to Word Sense Disambiguation. Pitman, London, U.K.

CRESTAN, E., EL-BZE, M., AND LOUPY, C. D. 2001. Improving WSD with multi-level view of context monitored by similarity measure. In Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (Senseval-2, Toulouse, France). 67–70.

CRUSE, D. A. 1986. Lexical Semantics. Cambridge University Press, Cambridge, U.K.

CRUSE, D. A.2000. Meaning in Language: An Introduction to Semantics and Pramatics. Oxford University Press, Oxford, New York.

CUADROS, M. AND RIGAU, G. 2006. Quality assessment of large scale knowledge resources. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP, Sydney, Australia). 534–541.

CUCCHIARELLI, A. AND VELARDI, P. 1998. Finding a domain-appropriate sense inventory for semantically tagging a corpus. J. Nat. Lang. Eng. 4, 4, 325–344.

DAELEMANS, W., VAN DEN BOSCH, A., AND ZAVREL, J. 1999. Forgetting exceptions is harmful in language learning. Mach. Learn. 34, 1, 11–41.

DAHLGREN, K.G. 1988. Naive Semanticsfor Natural Language Understanding. Kluwer Academic Publishers, Boston.

DAGAN, I., MARCUS, S, and MARKOVITCH S. 1993. Contextual wordsimilarity and estimation from sparse data. In Proceedings of the 31st AnnualMeeting, Columbus, OH, June. Association for Computational Linguistics.

DAGAN, I. AND ENGELSON, S. 1995. Selective sampling in natural language learning. In Proceedings of the IJCAI Workshop on New Approaches to Learning for Natural Language Processing (Montreal, P.Q., Canada). 41–48.

DAGAN, I. AND ITAI, A. 1994. Word sense disambiguation using a second language monolingual corpus. Computat. Ling. 20, 4, 563–596.

DECADT, B., HOSTE, V., DAELEMANS, W., AND VAN DEN BOSCH, A. 2004. GAMBL, genetic algorithm optimization of memory-based WSD. In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain). 108–112.

DE GROOT, A.M.B. 1983. The range ofautomatic spreading activation in word priming. Journal of Verbal learning and Verbal Behavior, 22(4):417-436.

DEMPSTER, A., LAIRD, N., AND RUBIN, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

DIAB, M. 2003. Word sense disambiguation within a multilingual framework. Ph.D. dissertation. University of Maryland, College Park, College of Park, MD.

DILL, S., EIRON, N., GIBSON, D., GRUHL, D., GUHA, R., JHINGRAN, A., KANUNGO, T., RAJAGOPALAN, S., TOMKINS, A., TOMLINE, J. A., AND ZIEN, J.Y. 2003. Semtag and Seeker: Bootstrapping the semantic Web via automated semantic annotation. In Proceedings of the 20th International Conference on World Wide Web (WWW, Budapest, Hungary). 178–186.

DIXON, R.M.W. 1971. “A Method of Semantic Description”. In: Steinberg, D.D. and Jakobovits, L.A. (ed.). Semantics. Cambridge University Press, New York

DOLAN, W. B. 1994. Word sense ambiguation: Clustering related senses. In Proceedings of the 15th Conference on Computational Linguistics (COLING). Kyoto, Japan.

DOROW, B. AND WIDDOWS, D. 2003. Discovering corpus-specific word senses. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (Budapest, Hungary). 79–82.

DUFFIELD,C. J.,HWANG, J. D.,BROWN, S.W.,DLIGACH,D.,VIEWEG, S. E.,DAVIS, J., AND PALMER,M. 2007. Criteria for the manual grouping of verb senses. In Proceedings of the Linguistic Annotation Workshop (Prague, Czech Republic).

EARL, L.L. 1973. Use of word governmentin resolving syntactic and semantic ambiguities. Information Storage and Retrieval, 9:639-664.

EDMONDS, P. 2000. Designing a task for SENSEVAL-2. Tech. note. University of Brighton, Brighton. U.K.

EDMONDS, P. AND COTTON, S. 2001. Senseval-2: Overview. In Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (Senseval-2, Toulouse, France). 1–6.

EDMONDS, P. AND KILGARRIFF, A. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. J. Nat. Lang. Eng. 8, 4, 279–291.

ESCUDERO, G., M`ARQUEZ, L., AND RIGAU, G. 2000a. Boosting applied to word sense disambiguation. In Proceedings of the 11th European Conference on Machine Learning (ECML, Barcelona, Spain). 129–141.

ESCUDERO, G., M`ARQUEZ, L., AND RIGAU, G. 2000b. Naive Bayes and exemplar-based approaches to word sense disambiguation revisited. In Proceedings of the 14th European Conference on Artificial Intelligence (ECAI, Berlin, Germany). 421–425.

ESCUDERO, G., M`ARQUEZ, L., AND RIGAU, G. 2000c. On the portability and tuning of supervised word sense disambiguation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC, Hong Kong, China). 172–180.

ETZIONI, O., CAFARELLA, M., DOWNEY, D., KOK, S., POPESCU, A.-M., SHAKED, T., SODERLAND, S., WELD, D. S., AND YATES,A. 2004. Web-scale information extraction in Knowitall. In Proceedings of the 13th International Conference on World Wide Web (WWW, New York, NY). 100–110.

FELDMAN, J.A. and Ballard D.H.1982. Connectionist models and their properties. Cognitive Science, 6(3):205-254.

FELLBAUM, C., Ed. 1998. WordNet: An Electronic Database. MIT Press, Cambridge, MA: MIT Press.

FELLBAUM, C., PALMER, M., DANG, H. T., DELFS, L., ANDWOLF, S. 2001. Manual and automatic semantic annotation with WordNet. In Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (Pittsburgh, PA). 3–10.

FILLMORE, C.J. and ATKINS B.T.S..1991. Invited lecture. Presented at the 29th Annual Meeting of the Associationfor Computational Linguistics, Berkeley, CA, June.

FISCHLER, I. 1977. Semantic facilitationwithout association in a lexical decision task. Memory and Cognition, 5(3):335-339.

FLORIAN, R., CUCERZAN, S., SCHAFER, C., AND YAROWSKY, D. 2002. Combining classifiers for word sense disambiguation. J. Nat. Lang. Eng. 8, 4, 1–14.

FREUND, Y. AND SCHAPIRE, R. 1999. A short introduction to boosting. J. Japanese Soci. Artific. Intell. 14, 771–780.

FU, K. 1982. Syntactic Pattern Recognition and Applications. Prentice-Hall, Engelwood Cliffs, NJ.

FUJII, A., INUI, K., TOKUNAGA, T., AND TANAKA, H. 1998. Selective sampling for example-based word sense disambiguation. Computat. Ling. 24, 4, 573–598.

GALE, W. A., CHURCH, K., AND YAROWSKY, D. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (Newark, NJ). 249–256.

GALE, W. A., CHURCH, K., AND YAROWSKY, D. 1992b. A method for disambiguating word senses in a corpus. Comput. Human. 26, 415–439.

GALE, W.A., CHURCH, K.W., AND YAROWSKY, D. 1992c. One sense per discourse. In Proceedings of the DARPA Speech and Natural Language Workshop (Harriman, NY). 233–237.

GALE, W. A., CHURCH, K.W., AND YAROWSKY, D. 1992d. Using bilingual materials to develop word sense disambiguation methods. In Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (Montreal, P.Q., Canada). 101–112. Brown, Della Pietra, Mercer(1991).

GALE, W. A., KENNETH W. CHURCH, AND DAVID YAROWSKY. 1992e. Work on statistical methods for word sense disambiguation. In Robert Goldman Peter Norvig, Eugene Charniak and Bill Gale (eds.), *Working Notes of AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pp. 54-60, Menlo Park, CA. AAAI Press.

GALE, W. A., Church, K.W. and Yarowsky D. 1993. A method for disambiguating word senses in a largecorpus. Computers and the Humanities, 26:415-439.

GALLEY, M. AND MCKEOWN, K. 2003. Improving word sense disambiguation in lexical chaining. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico). 1486–1488.

GIRJU, R.,BADULESCU, A., ANDMOLDOVAN,D. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (Edmonton, Alta., Canada). 1–8.

GLIOZZO, A., MAGNINI, B., AND STRAPPARAVA, C. 2004. Unsupervised domain relevance estimation for word sense disambiguation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP, Barcelona, Spain). 380–387.

GOLUB, G. H. AND VAN LOAN, C. F. 1989. Matrix Computations. The John Hopkins University Press, Baltimore, MD.

GOOD, I.J. 1953. The population frequencies of species and the distribution of population parameters. Biometrika, 40(3/4):237-264.

GRAFF, D. 2003. English gigaword. Tech. rep. LDC2003T05. Linguistic Data Consortium, Philadelphia, PA.

GRIMSHAW, J. 1990. Argument Structure. Cambridge: MIT Press.

GRISHMAN, R., Catherine MacLeod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. InProceedings of the 15th International Conference on Computational Linguistics, COLING'94, pages 268-272, Kyoto, Japan, August.

GROZEA, C. 2004. Finding optimal parameter settings for high performance word sense disambiguation. In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain). 125–128.

GRUBER, J. 1976. Lexical Structures in Syntax and Semantics. New York: North Holland.

GRUBER, T. R. 1993. Toward principles for the design of ontologies used for knowledge sharing. In Proceedings of the International Workshop on Formal Ontology (Padova, Italy).

HALLIDAY, M. A. K. 1961. Categories of thetheory of grammar. Word, 17:241-292.

HALLIDAY, M. A. K. 1966. Lexis as a linguisticMemory of J. R. Firth. Longman, London, pages 148-163.

HALLIDAY, M. A. AND HASAN, R., Eds. 1976. Cohesion in English. Longman Group Ltd, London, U.K.

HARABAGIU, S., MILLER, G., AND MOLDOVAN, D. 1999. WordNet 2—a morphologically and semantically enhanced resource. In Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources. 1–8.

HARRIS, Z.S. 1954. Distributionalstructure. Word, 10:146-162.

HASSAN, H., HASSAN, A., AND NOEMAN, S. 2006. Graph based semi-supervised approach for information extraction. In Proceedings of the TextGraphs Workshop in the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL, New York, NY). 9–16.

HAWKINS, P. AND NETTLETON, D. 2000. Large scale WSD using learning applied to senseval. Comput. Human. 34, 1-2, 135–140.

HAYES, P.J. 1976. A Process to implement some word-sense disambiguation. Working paper 23. Institut pour les Etudes Semantiques et Cognitives, Université de Genéve.

HAYES, P.J. 1977a. On Semantic Nets, Frames and Associations. In Proceedings of the 5th International Joint Conference on Artificial Intelligence, pages 99-107.

HAYES, P.J. 1977b. Some Association-based Techniques for Lexical Disambiguation by Machine. Doctoral dissertation, Départment de Mathématiques, Ecole Polytechnique Fédérale de Lausanne.

HAYES, P.J. 1978. Mapping input into chemas. Technical Report 29, Department of Computer Science, University of Rochester.

HEARST, M. 1991. Noun homograph disambiguation using local context in large text corpora. In Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora (Oxford, U.K). 1–19.

HEARST, M. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics (COLING, Nantes, France). 539–545.

HINDLE, D. AND ROOTH,M. 1993. Structural ambiguity and lexical relations. Computat. Ling. 19, 1, 103–120.

HIRST,G. (Ed). 1987. Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press, Cambridge, U.K.

HIRST, G. AND ST-ONGE,D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In WordNet: An Electronic Lexical Database, C. Fellbaum, Ed. MIT Press, Cambridge, MA, 305–332.

HJEMSLEV, L. 1953. Prolegomena to aTheory of Language. Translated from Danish. Indiana University, Bloomington,IN.

HOSTE, V., HENDRICKX, I., DAELEMANS, W., AND VAN DEN BOSCH, A. 2002. Parameter optimization for machine learning of word sense disambiguation. J. Nat. Lang. Eng. 8, 4, 311–325.

HOVY, E., MARCUS, M., PALMER, M., RAMSHAW, L., AND WEISCHEDEL, R. 2006. Ontonotes: The 90% solution. In Proceedings of the Human Language Technology Conference of the NAACL, Comp. Volume (New York, NY). 57–60.

IDE, N. 2000. Cross-lingual sense determination: Can it work? Comput. Human. 34, 1–2, 223–234.

IDE, N. 2006. Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Computational Linguistics and Intelligent Text, A. Gelbukh, Ed. Lecture Notes in Computer Science, vol. 3878. Springer, Berlin, Germeny, 13–27.

IDE, N., ERJAVEC, T., AND TUFIS, D. 2001. Automatic sense tagging using parallel corpora. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (Tokyo, Japan). 83–89.

IDE, N., ERJAVEC, T., AND TUFIS, D. 2002. Sense discrimination with parallel corpora. In Proceedings of ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions (Philadelphia, PA). 54–60.

ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date: February 2009.

IDE, N. AND SUDERMAN, K. 2006. Integrating linguistic resources: The American National Corpus model. In Proceedings of the 5th Language Resources and Evaluation Conference (LREC, Genoa, Italy).

IDE, N. AND V´ERONIS, J. 1993. Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In Proceedings of the Workshop on Knowledge Bases and Knowledge Structures (Tokyo, Japan). 257–266.

IDE, N. AND V´ERONIS, J. 1998. Word sense disambiguation: The state of the art. Computat. Ling. 24, 1, 1–40. IDE, N. AND WILKS, Y. 2006. Making sense about sense. In Word Sense

Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 47–73.

JACQUEMIN, B., BRUN, C., AND ROUX, C. 2002. Enriching a text by semantic disambiguation for information extraction. In Proceedings of the Workshop on Using Semantics for Information Retrieval and Filtering in the 3rd International Conference on Language Resources and Evaluations (LREC, Las Palmas, Spain).

JIANG, J. J. AND CONRATH, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics (Taiwan, ROC).

JOACHIMS, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning (ECML, Heidelberg, Germany). 137–142.

JOHNSTON, M. AND BUSA, F. 1996. Qualia structure and the compositional interpretation of compounds. In Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons (Santa Cruz, CA).

JORGENSEN, J. 1990. The psychological reality of word senses. Journal of Psycholinguistic Research, 19:167-190.

JURAFSKY, D. ANDMARTIN, J. 2000. Speech and Language Processing. Prentice Hall, Upper Saddle River, NJ.

KAPLAN, A. 1950. An Experimental Study of Ambiguity and Context. Mimeographed, (Publihsed in 1955 in Machnaical Translation, 2(2):39-46.

KATZ, J.J. AND FODOR, J.A. 1963. The Structure of a Semantic Theory. Language 39.170-210

KATZ, J.J. AND FODOR, J.A. 1990. Women, Fire and Dangerous Things. Chicago: University of Chicago Press.

KELLY, E. AND STONE, P. 1975. Computer Recognition of English Word Senses. Vol. 3 of North Holland Linguistics Series. Elsevier, Amsterdam, The Netherlands.

KEOK, L. Y. AND NG, H. T. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP, Philadelphia, PA). 41–48.

KILGARRIFF, A. 1992. Polysemy. Ph.D.thesis. University of Sussex, UK.

KILGARRIFF, A. 1993. Dictionary word sense distinctions: An enquiry into theirnature. Computers and the Humanities, 26:365-387.

KILGARRIFF, A. 1997. I don't believe in word senses. Comput. Human. 31, 2, 91–113.

KILGARRIFF, A. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC, Granada, Spain). 1255–1258.

KILGARRIFF, A. 2006. Word senses. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 29–46.

KILGARRIFF, A. ANDGREFENSTETTE,G. 2003. Introduction to the special issue on theWeb as corpus. Computat. Ling. 29, 3, 333–347.

KILGARRIFF, A AND DAVID, T. 2001. WASP-Bench: An MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. In Proceedings of MT Summit VII, pp. 187-190. Santiago de Compostela.Navigli, R. 2009.

KILGARRIFF, A. AND PALMER,M. 2000. Introduction to the special issue on Senseval. Comput. Human. 34, 1-2, 1–13.

KILGARRIFF, A. AND ROSENZWEIG, J. 2000. English Senseval: Report and results. In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC, Athens, Greece).

KILGARRIFF, A., RYCHLY, P., SMRZ, P., AND TUGWELL, D. 2004. The Sketch Engine. In Proceedings of the 11th EURALEX International Congress (Lorient, France). 105–116.

KILGARRIFF, A. AND YALLOP, C. 2000. What's in a thesaurus? In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC, Athens, Greece). 1371–1379.

KINTSCH, W and MROSS E.E. 1985. Context effects in word identification. Journal of Memory and Language,24(3):336-349.

KLEIN, D., TOUTANOVA, K., ILHAN, T. H., KAMVAR, S. D., AND MANNING, C. D. 2002. Combining heterogeneous classifiers for word-sense disambiguation. In Proceedings of the ACL workshop on Word Sense Disambiguation: Recent Successes and Future Directions (Philadelphia, PA). 74–80.

KOHOMBAN, U. S. AND LEE, W. S. 2005. Learning semantic classes for word sense disambiguation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (Ann Arbor, MI). 34–41.

KROVETZ, R. AND CROFT, W. B. 1992. Lexical ambiguity and information retrieval. ACM Trans. Inform. Syst. 10, 2, 115–141.

KUCERA, H. AND FRANCIS, W. N. 1967. Computational Analysis of Present-Day American English. Brown University Press, Providence, RI.

KUDO, T. AND MATSUMOTO, Y. 2001. Chunking with support vector machines. In Proceedings of NAACL (Pittsburgh, PA). 137–142.

LAPATA, M. AND KELLER, F. 2007. An information retrieval approach to sense ranking. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL, Rochester, NY). 348–355.

LEACOCK, C. AND CHODOROW, M. 1998. Combining local context andWordNet similarity for word sense identification. In WordNet: An electronic Lexical Database, C. Fellbaum, Ed. MIT Press, Cambridge, MA, 265–283.

LEACOCK, C., CHODOROW, M., AND MILLER, G. 1998. Using corpus statistics and WordNet relations for sense identification. Computat. Ling. 24, 1, 147–166.

LEACOCK, C., TOWELL, G., AND VOORHEES, E. 1993. Corpus-based statistical sense resolution. In Proceedings of the ARPA Workshop on Human Language Technology (Princeton, NJ). 260–265.

LEE, L. 1999. Measures of distributional similarity. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (College Park, MD). 25–32.

LEECH, G.N. 1974. Semantics. Penguin, Harmondsworth.

LEECH, G.N. 1983. Semantics and Cognition. MIT Press. Cambridge MA.

LEECH, G.N. 1990.Semantic Structures. MIT Press, Cambridge, MA.

LEHERER, A. 1974. Semantic Fields and Lexical Structure. North Holland Publishing Company, London.

LENAT, D.B. and Guha, R.V. 1990. Building Large Knowledge-based Systems. Addison-Wesley, Reading, MA.

LESK, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th SIGDOC (New York, NY). 24–26.

LEVIN, B, 1991. Building a Lexicon: The Contribution of Linguistic Theory. International Journal of Lexicography 4.205-226.

LI, H. AND ABE, N. 1998. Generalizing case frames using a thesaurus and the MDL principle. Computat. Ling. 24, 2, 217–244.

LIDDY, E.D. and PAIK, W. 1993. Statistically-guided word sense disambiguation. In Proceedings of the AAAI Fall Symposium Series, pages 98-107.

LIN,D. 1998a. Automatic retrieval and clustering of similar words. In Proceedings of the 17th International Conference on Computational linguistics (COLING, Montreal, P.Q., Canada). 768–774.

LIN, D. 1998b. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning (ICML, Madison, WI). 296–304.

LIN, D. AND PANTEL, P. 2002. Discovering word senses from text. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edmonton, Alta., Canada). 613–619.

LITKOWSKI,K. C. 1978. Models of the semantic structure of dictionaries. Amer. J. Computat. Ling. 81, 25–74.

LITKOWSKI, K. C. 2005. Computational lexicons and dictionaries. In Encyclopedia of Language and Linguistics (2nd ed.), K. R. Brown, Ed. Elsevier Publishers, Oxford, U.K., 753–761.

LITKOWSKI, K. C. AND HARGRAVES, O. 2007. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 24–29.

LUGER, G. F. 2004. Artificial Intelligence: Structures and Strategies for Complex Problem-Solving, 5th ed. Addison Wesley, Reading, MA.

LUK, A.K. 1995. Statistical sensedisambiguation with relatively small corpora using dictionary definitions. In Proceedings of the 33rd Annual Meeting, pages 181-188. Cambridge, MA.Association for Computational Linguistics.

LUPKER, S.J. 1984. Semantic priming without association: A second look. Journal of Verbal Learning and Verbal Behavior, 23(6):709-733.

LYONS, J. 1963. Structural Semantics. Blackwell, Oxford

LYONS, J. 1975. Semantics, 2 volumes. Cambridge University Press, New York.

LYONS, J. 1995. Linguistic Semantics: An Introduction. Cambridge University Press. Cambridge:

LYONS, J. 1977. Semantics. Cambridge University Press, Cambridge, UK.

MACLEOD, C., Ralph Grishman, and Adam Meyers. Forthcoming. A large syntactic dictionary for natural language processing. Computers and the Humanities.

MADHU, S. AND LYTLE, D.W. 1965. A Figure of Merit Technique for the Resolution of Non-grammatical Ambiguity. Mechanical Translation, 8(2):9-13.

MAGNINI, B. AND CAVAGLI`A, G. 2000. Integrating subject field codes into WordNet. In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC, Athens, Greece). 1413–1418.

MALIN, B., AIROLDI, E., AND CARLEY, K. M. 2005. A network analysis model for disambiguation of names in lists. Computat. Math. Organizat. Theo. 11, 2, 119–139.

MALLERY, J. C. 1988. Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. Ph.D. dissertation. MIT Political Science Department, Cambridge, MA.

MANNING, C. AND SCH¨UTZE, H. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

MARKERT, K. AND NISSIM, M. 2007. Semeval-2007 task 08: Metonymy resolution at Semeval-2007. In Proceedings of the 4th InternationalWorkshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 36–41.

MALAKHOVSKI, L. V. 1987. Homonyms inEnglish dictionaries. In R. W. Burchfield, editor, Studies in Lexicography. OxfordUniversity Press, Oxford, UK, pages 36-51.

MARKOWITZ, J., Thomas Ahlswede, and Martha Evens. 1986. Semantically significant patterns in dictionary definitions. In Proceedings of the 24th Annual Meeting, pages 112-119, Association for Computational Linguistics.

MARTINET, A. 1960. Eléments de linguistique générale. Armand Colin, Paris.

M`ARQUEZ, L., ESCUDERO, G., MART`INEZ, D., AND RIGAU, G. 2006. Supervised corpus-based methods for WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 167–216.

MARTINEZ, D. 2004. Supervised word sense disambiguation: Facing current challenges, Ph.D. dissertation. University of the Basque Country, Spain.

MASTERMAN, M. 1957. The Thesaurus in Syntax and Seamtnics. Mechanical Translation, 4:1-2.

MASTERMAN, M. 1961. Semantic message detection for machine translation, using intelingua. In 1961 International Conference on Machine Translation of Langauges and Applied Languague Analysis, Her Majesty's Stationery Office, London, pages 437-475.

MCCARTHY, D. 2006. Relating wordnet senses for word sense disambiguation. In Proceedings of the ACL Workshop on Making Sense of Sense (Trento, Italy). 17–24.

MCCARTHY, D. AND CARROLL, J. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. Computat. Ling. 29, 4, 639–654.

MCCARTHY, D., KOELING, R., WEEDS, J., AND CARROLL, J. 2004. Finding predominant senses in untagged text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (Barcelona, Spain). 280–287.

MCCARTHY, D., KOELING, R.,WEEDS, J., AND CARROLL, J. 2007. Unsupervised acquisition of predominant word senses. Computat. Ling. 33, 4, 553–590.

MCCARTHY, D. ANDNAVIGLI, R. 2007. Semeval-2007 task 10: English lexical substitution task. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 48–53.

McCLELLAND, J.L. and RUMELHART D.E.. 1981. An interactive activation of context effects in letter perception:Part 1. An account of basic findings. Psychological Review, 88:375-407.

McCRAY, A. T. AND NELSON, S. J. 1995. The representation of meaning in the UMLS. Meth. Inform. Med. 34, 193–201.

MCCULLOCH, W. AND PITTS, W. 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133.

McROY, S.W. 1992. Using multiple knowledge sources for word sense discrimination. Computational Linguistics, 18(1):1-30.

MEILLET, A. 1926. Linguistique historiqueet linguistique g~ndrale. Volume 1. Second edition. Champion, Paris.

MEYER, D.E. and Schvaneveldt, R.W. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90(2):227-234.

MICHIELS, A. 1982. Exploiting a Large Dictionary Data Base. Ph.D. thesis,Universit4 de Liege, Liege, Belgium.

MIHALCEA, R. 2002a. Bootstrapping large sense tagged corpora. In Proceedings of the 3rd International Conference on Language Resources and Evaluations (LREC, Las Palmas, Spain).

MIHALCEA, R. 2002b. Word sense disambiguation with pattern learning and automatic feature selection. J. Nat. Lang. Eng. 8, 4, 348–358.

MIHALCEA, R. 2004. Co-training and self-training for word sense disambiguation. In Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL, Boston, MA). 33–40.

MIHALCEA, R. 2006. Knowledge-based methods for WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, 107–131.

MIHALCEA, R. AND EDMONDS, P., Eds. 2004. Proceedings of the 3rd InternationalWorkshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain).

MIHALCEA, R. AND FARUQUE, E. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain). 155–158.

MIHALCEA, R. AND MOLDOVAN, D. 1999. An automatic method for generating sense tagged corpora. In Proceedings of the 16th National Conference on Artificial Intelligence (AAAI, Orlando, FL). 461–466.

MIHALCEA, R., TARAU, P., AND FIGA, E. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In Proceedings of the 20th International Conference on Computational Linguistics (COLING, Geneva, Switzerland). 1126–1132.

MILLER, G.A.1990. “Nouns in WordNet:a lexical inheritance system”. International Journal of Lexicography Vol 3, No. 4, 245-264.

MILLER, G.A. 1991. “Science of Words”. New York: Scientific American Library.

MILLER, G. A., BECKWITH, R., FELLBAUM, C. D., GROSS, D., AND MILLER, K. 1990. WordNet: An online lexical database. Int. J. Lexicography. 3, 4, 235–244.

MILLER, G. A., LEACOCK, C., TENGI, R., AND BUNKER, R. T. 1993. A semantic concordance. In Proceedings of the ARPA Workshop on Human Language Technology. 303–308.

MITCHELL, T. 1997. Machine Learning. McGraw Hill, New York, NY .

MOHAMMAD, S. AND HIRST, G. 2006. Determining word sense dominance using a thesaurus. In Proceedings of the 11th Conference on European chapter of the Association for Computational Linguistics (EACL, Trento, Italy). 121–128.

MOONEY, R. J. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP). 82–91.

Morris, Jane. 1988. Lexical cohesion, thethesaurus, and the structure of text. Technical Report CSRI 219, ComputerSystems Research Institute, University of Toronto, Toronto, Canada.

MORRIS, J. and HIRST G. 1991.Lexical cohesion computed by thesaural relations as an indicator of the structure oftext. Computational Linguistics, 17(1):21-48.

MURATA, M.,UTIYAMA, M.,UCHIMOTO, K.,MA, Q., AND ISAHARA,H. 2001. Japanese word sense disambiguation using the simple Bayes and support vector machine methods. In Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (Senseval-2, Toulouse, France). 135–138.

NAVIGLI, R. 2005. Semi-automatic extension of large-scale linguistic knowledge bases. In Proceedings of the 18th Florida Artificial Intelligence Research Society Conference (FLAIRS, Clearwater Beach, FL). 548–553.

NAVIGLI, R. 2006a. Consistent validation of manual and automatic sense annotations with the aid of semantic graphs. Computat. Ling. 32, 2, 273–281.

NAVIGLI, R. 2006b. Experiments on the validation of sense annotations assisted by lexical chains. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL, Trento, Italy). 129–136.

NAVIGLI, R. 2006c. Meaningful clustering of senses helps boost word sense disambiguation performance. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL, Sydney, Australia). 105–112.

NAVIGLI, R. 2008. A structural approach to the automatic adjudication of word sense disagreements. J. Nat. Lang. Eng. 14, 4, 547–573.

NAVIGLI, R. 2009. Word Sense Disambiguation: A Survey. ACM Computing Surveys, Vol 41, No.2, Article 10, Publication date February 2009.

NAVIGLI, R. AND LAPATA, M. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI, Hyderabad, India). 1683–1688.

NAVIGLI, R., LITKOWSKI, K. C., AND HARGRAVES, O. 2007. Semeval-2007 task 07: Coarse-grained English allwords task. In Proceedings of the 4th InternationalWorkshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 30–35.

NAVIGLI, R. AND VELARDI, P. 2004. Learning domain ontologies from document warehouses and dedicated Websites. Computat. Ling. 30, 2, 151–179.

ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date: February 2009

NAVIGLI, R. AND VELARDI, P. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Trans. Patt. Anal. Mach. Intell. 27, 7, 1075–1088.

NAVIGLI, R., VELARDI, P., AND GANGEMI, A. 2003. Ontology learning and its application to automated terminology translation. IEEE Intell. Syst. 18, 1, 22–31.

NG, H. T. AND LEE, H. B. 1996. Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (Santa Cruz, CA). 40–47.

NG, H. T., WANG, B., AND CHAN, Y. S. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (Sapporo, Japan). 455–462.

NG, T. H. 1997. Getting serious about word sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington D.C.). 1–7.

NG, V. AND CARDIE, C. 2003. Weakly supervised natural language learning without redundant views. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL, Edmonton, Alta., Canada). 173–180.

NIDA, E.A. 1975a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton

NIDA, E.A. 1975.b. Exploring Semantic Structure. The Hague: Mouton

NIU, C., LI, W., SRIHARI, R., AND LI, H. 2005. Word independent context pair classification model for word sense disambiguation. In Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL, Ann Arbor, MI).

NIU, Z.-Y., JI, D.-H., AND TAN, C.-L. 2007. I2r: Three systems for word sense discrimination, Chinese word sense disambiguation, and English word sense disambiguation. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 177–182.

PALMER, M., BABKO-MALAYA, O., AND DANG, H. T. 2004. Different sense granularities for different applications. In Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems in HLT/NAACL (Boston, MA). 49–56.

PALMER, M., DANG, H., AND FELLBAUM, C. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. J. Nat. Lang. Eng. 13, 2, 137–163.

PALMER, M., NG, H. T., AND DANG, H. T. 2006. Evaluation of WSD systems. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 75–106.

PANTEL, P. 2005. Inducing ontological co-occurrence vectors. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (Ann Arbor, MI). 125–132.

PATRICK, A.B. 1985. An Exploration of Abstract Thesaurus Instantiation. M. Sc. thesis, University of Kansas, Lawrence,KS.

PEASE, A.,NILES, I., AND LI, J. 2002. The suggested upper merged ontology: A large ontology for the semantic Web and its applications. In Proceedings of the AAAI-2002 Workshop on Ontologies and the Semantic Web (Edmonton, Alta., Canada).

PEDERSEN, T. 1998. Learning probabilistic models of word sense disambiguation, Ph.D. dissertation. Southern Methodist University, Dallas, TX.

PEDERSEN, T. 2006. Unsupervised corpus-based methods for WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 133–166.

PEDERSEN, T., BANERJEE, S., AND PATWARDHAN, S. 2005. Maximizing semantic relatedness to perform word sense disambiguation. Res. rep. UMSI 2005/25. University of Minnesota Supercomputing Institute, Minneapolis, MN.

PEDERSEN, T. AND BRUCE, R. 1997. Distinguishing word senses in untagged text. In Proceedings of the 1997 Conference on Empirical Methods in Natural Language Processing (EMNLP, Providence, RI). 197–207.

PEDERSEN, T., PATWARDHAN, S., AND MICHELIZZI, J. 2004. WordNet::Similarity—measuring the relatedness ofconcepts. In Proceedings of the 19th National Conference on Artificial Intelligence (AAAI, San Jose, CA) 144–152.

PENNACCHIOTTI, M. AND PANTEL, P. 2006. Ontologizing semantic relations. In Proceedings of the 44th Association for Computational Linguistics (ACL) Conference joint with the 21th Conference on Computational Linguistics (COLING, Sydney, Australia). 793–800.

PEREIRA, F. and Tishby N. 1992. Distributional similarity, phase transitions and hierarchical clustering. Working Notes of the AAAI Symposium on Probabilistic Approaches to Natural Language, pages 108-112, Cambridge, MA, October.

PEREIRA, F., TISHBY, N., AND LEE, L. 1993. Distributional clustering of English. In Proceeedings of the 31st Annual Meeting, pages 183-190, Ohio State University, Colombus, OH, June.Association for Computational Linguistics.

PHILPOT, A., HOVY, E., AND PANTEL, P. 2005. The Omega Ontology. In Proceedings of the IJCNLP Workshop on Ontologies and Lexical Resources (OntoLex, Jeju Island, South Korea). 59–66.

PIANTA, E., BENTIVOGLI, L., AND GIRARDI, C. 2002. MultiWordNet: Developing an aligned multilingual database. In Proceedings of the 1st International Conference on Global WordNet (Mysore, India) 21–25.

PRADHAN, S., LOPER, E., DLIGACH, D., AND PALMER, M. 2007. Semeval-2007 task-17: English lexical sample, SRL and all words. In Proceedings of the 4th InternationalWorkshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 87–92.

PROCTOR, P., Ed. 1978. Longman Dictionary of Contemporary English. Longman Group, Harlow, U.K.

PURANDARE, A. AND PEDERSEN, T. 2004. Improving word sense discrimination with gloss augmented feature vectors. In Proceedings of theWorkshop on Lexical Resources for theWeb andWord Sense Disambiguation (Puebla, Mexico). 123–130.

PUSTEJOVSKY, J. 1991. The generative lexicon. Computat. Ling. 17, 4, 409–441.

PUSTEJOVSKY, J. 1995. The Generative Lexicon. MIT Press, Cambridge, MA.

QUILLIAN, M. R. 1962a. A Revised Design for an Understanding Machine. Mechanical Translation, 7(1):17-29.

QUILLIAN, M. R.1962b. A Semantic Coding Technique for Mechanical English paraphrasing International Memorandum of the Mechanical Translation Group, Rearch Laboratory of Electronics, MIT, August.

QUILLIAN, M. R.1967. Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities. Behavioral Science, 12:410-430.

QUILLIAN, M. R. 1968. Semantic Memory. In M. Minsky (ed). Semantic Information Processing. MIT Press, Cambridge, MA, pages 227-270.

QUILLIAN, M. R. 1969. The Teachable Langauge Comprehender: A simulation program and theory of Language. Communications of the ACM, 12(8):459-476.

QUINLAN, J. R. 1986. Induction of decision trees. Mach. Learn. 1, 1, 81–106.

QUINLAN, J. R. 1993. Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA.

QUINE, W.V. 1960. Word and Object. MIT Press, Cambridge, MA.

RADA, R., MILI, H., BICKNELL, E., AND BLETTNER, M. 1989. Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybernet. 19, 1, 17–30.

RAJENDRAN, S. 1978. Syntax and Semantics of Tamil Verbs. Ph.D. Thesis. Poona: University of Poona.

RAJENDRAN, S. 1982. Verbs of 'seeing' in Tamil. Poona: Bulletin of the Deccan College Research Institute, Poona 41:151-159.

RAJENDRAN, S. 1983. Semantics of Tamil Vocabulary. (Report of the UGC sponsored Postdoctoral Work in manuscript). Poona: Deccan College Post-Doctoral Research Institute.

RAJENDRAN, S. 1983. Temporal Expression in Tamil. Bulletin of the Deccan College: Poona 42:138-147.

RAJENDRAN, S. 1995. Componential Analysis of 'eating' in Tamil. PILC Journal of Dravidian Studies, 5.2:175-181.

RAJENDRAN, S. 1995. Towards a Compilation of a Thesaurus for Modern Tamil. South Asian Language Review 5.1:62-99.

RAJENDRAN, S. 2000. A Model for a 'Theoretical Dictionary of Tamil'. Tamil Civilization vol. 14-18, March 1996-2000, 99-105.

RAJENDRAN, S. 2002. Semantic Structure of Directonal Verbs of Movement in Tamil. Language in India 2:5, www.languageininda.com.

RAJENDRAN, S. 2002. "Semantic Structure of Verbs of Transfer in Tamil". Language in India 2:8, www.langugeinindia.com.

RAJENDRAN, S. 2002. Preliminaries to the preparation of wordnet for Tamil. Language in India 2:1, www.languageinindia.com

RAJENDRAN, S. 2003. Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages. Language in India 3:1, www.languageinindia.com

RAJENDRAN, S. 2003. Dravidian WordNet: A Proposal. R.M.Sundaram et al (edited), Facets of Language, Thanjavur. 467-497.

RAJENDRAN, S. 2003. Creating Generative Lexicon from MRDs: Tamil Experience. Rajeev Sangal et al (edited) Recent Advances in Natural Language Processing: Proceedings of the International Conference of Natural Language Processing (ICON 2003). IIIT, Hyderabad, NCST, Bombay & CIIL, Mysore, 83-91

RAJENDRAN, S. 2006. Languge Technology in Tamil. Language in India 6:8, www.languageinindia.com, 2006.

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 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. "Resolution of lexical ambiguity in Tamil." Languge in Inda, vol. 14:1, January 2014, 2014 ISSN 1930-2940

RAJENDRAN S, ARULMOZI S, KUMARA SHANMUGAM B, BASKARAN S AND THIAGARAJAN S. 2002. Tamil WordNet. In Proceedings of the First Global WordNet Conference, CIIL, Mysore, 2002, 271-274.

RAJENDRAN, S. AND BASKARAN, S. 2002. Preparation of Electronic Thesaurus for Tamil. (Co-author: S.Baskaran). In Proceedings of the International Conference on Natural Language Processing. NCST, Mumbai, 2002

RAJENDRAN S, SHIVAPRATAP G, DHANALAKSHMI V AND SOMAN K P. 2010. Building a WordNet for Dravidian Languages. In Proceedings of the 5th International Conference of the Global WordNet Association (GWA2010), Indian Institute of Technology, Mumbai, India

RAVICHANDRAN, D. AND HOVY, E. 2002. Learning surface text patterns for a question answering system. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (Philadelphia, PA). 41–47.

REIFLER, E. 1955. The Mechanical determination of meaning. In William N. Locke and A. Donald Booth (eds.) Machine Translation of Languages. John Wiley & Sons, New York, pages 136-164.

RESNIK, P. 1992. WordNet and distributional analysis: A class-based approach to statistical discovery. In Proceedings of the AAAI Workshop on Statistically-based Natural Language Processing Techniques, pages 48-56. San Jose, CA.

RESNIK, P. 1993a. Selection and Information: A Class-based Approach to Lexical Relationships. Ph.D. thesis, University of Pennsylvania. Also University of Pennsylvania Technical Report 93-42.

RESNIK, P. 1993b. Semantic classes and syntactic ambiguity. In Proceedings of the ARPA Workshop on Human Language Technology, pages 278-283.

RESNIK, P. 1995a. Disambiguating noun groupings with respect to WordNet senses. In Proceedings of the Third Workshop on Very Large Corpora, pages 54-68, Cambridge, MA.

RESNIK, P. 1995b. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95, pages 448-453, Montreal, Canada.

RESNIK, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI, Montreal, P.Q., Canada). 448–453.

RESNIK, P. 1997. Selectional preference and sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington, D.C.). 52–57.

RESNIK, P. 2006. Word sense disambiguation in NLP applications. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY.

RESNIK, P. AND SMITH, N. A. 2003. The Web as a parallel corpus. Computat. Ling. 29, 3, 349–380.

RESNIK, P. AND YAROWSKY, D. 1997. A perspective on word sense disambiguation methods and their evaluation.

In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington, D.C.). 79–86.

RESNIK, P. AND YAROWSKY,D. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. J. Nat. Lang. Eng. 5, 2, 113–133.

RESNIK, P. S., Ed. 1993. Selection and information: A class-based approach to lexical relationships, Ph.D. dissertation. University of Pennsylvania, Pennsylvania, Philadelphia, PA.

RICHARDSON, S. D., DOLAN, W. B., AND VANDERWENDE, L. 1998. Mindnet: Acquiring and structuring semantic information from text. In Proceedings of the 17th International Conference on Computational Linguistics (COLING, Montreal, P.Q., Canada). 1098–1102.

RICHENS, R. H. 1958. Intelingual machine translation.Computer Journal, 1(3):144-147.

RICHMOND, K., SMITH A., and AMITAY, E. 1997. Detecting Subject Boundaries Within Text: A Language Independent Statistical Approach. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, EMNLP-2, pages 47-54, Brown University, Providence, RI, August.

RIGAU, G., RODRIGUEZ, H., AND AGIRRE, E. 1998. Building accurate semantic taxonomies from monolingual MRDs. In Proceedings of the 17th International Conference on Computational Linguistics (COLING, Montreal, P.Q., Canada). 1103–1109.

RIVEST, R. L. 1987. Learning decision lists. Mach. Learn. 2, 3, 229–246.

ROBINS, R. H. 1987. Polysemy and the lexicographer. In R. W. Burchfield, editor, Studies in Lexicography. Oxford UniversityPress, Oxford, UK, pages 52-75.

ROGET, P. M. 1911. Roget's International Thesaurus, 1st ed. Cromwell, New York, NY.

RUSSELL, S. AND NORVIG, P. 2002. Artificial Intelligence: A Modern Approach, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

SALTON, G. 1968. Automatic Information Organization and Retrieval. McGraw-Hill, New York, NY.

SALTON, G. AND MCGILL, M. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY.

SANDERSON, M. 1994. Word sense disambiguation and information retrieval. In Proceedings of the Special Interest Group on Information Retrieval (SIGIR, Dublin, Ireland). 142–151.

SANDERSON, M. 2000. Retrieving with good sense. Inform. Retriev. 2, 1, 49–69.

SAVOVA, G., PEDERSEN, T., PURANDARE, A., AND KULKARNI, A. 2005. Resolving ambiguities in biomedical text with unsupervised clustering approaches. Res. rep. UMSI 2005/80. University of Minnesota Supercomputing Institute, Minneapolis, MN.

Schank, R.C. and Abelson R.P.1977. Scripts, Plans, Goals and Understanding. Lawrence Erlbaum, Hillsdale, NJ.

SCHAPIRE, R. E. AND SINGER, Y. 1999. Improved boosting algorithms using confidence-rated predictions. Mach. Learn. 37, 3, 297–336.

SCHÜTZE, H. 1992. Dimensions of meaning. In Supercomputing '92: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing. IEEE Computer Society Press, Los Alamitos, CA. 787–796.

SCHÜTZE, H. 1993. Word space. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, Advances in Neural Information Processing Systems 5. Morgan Kauffman, San Mateo, CA, pages 895-902.

SCHÜTZE, H. 1998. Automatic word sense discrimination. Computat. Ling. 24, 1, 97–124.

SCHÜTZE,H. AND PEDERSEN, J. 1995. Information retrieval based on word senses. In Proceedings of SDAIR'95 (Las Vegas, NV). 161–175.

SEDELOW, S.Y. and SEDELOW Jr. W.A. 1969. Categories and procedures for content analysis in the humanities. In George Gerbner, Ole Holsti, Klaus Krippendorf, William J. Paisley, and Philip J. Stone, editors, The Analysis of Communication Content. John Wiley & Sons, New York, pages 487--499.

SEDELOW, S.Y. and SEDELOW Jr. W.A. 1986. Thesaural knowledge representation. In Proceedings of the University of Waterloo Conference on Lexicology, pages 29--43, Waterloo, Canada.

SEDELOW, S.Y. and SEDELOW Jr. W.A. 1992. Recent model-based and model-related studies of a large-scale lexical resource (Roget's Thesaurus. In Proceedings of the 14th InternationalConference on Computational Linguistics, COLING'92, pages 1223-1227, Nantes,France, August.

Seidenberg, M.S., Michael K. Tanenhaus, James M. Leiman, and Marie A. Bienkowski. 1982. Automatic access of the meaning of ambiguous words in context:

Some limitations of knowledge-based processing. Cognitive Psychology, 14(4):489-537.

SILBER, H. G. AND MCCOY, K. F. 2003. Efficient text summarization using lexical chains. In Proceedings of the 5th International Conference on Intelligent User Interfaces (New Orleans, LA). 252–255.

Simpson, G.B. and Burgess, C. 1989. Implications of lexical ambiguity resolution for word recognition andcomprehension. In Steven Small, Garrison W. Cottrell, and Michael K. Tanenhaus,editors, Lexical Ambiguity Resolution: Perspectives from Psycholinguistics,Neuropsychology, and Arti~'cial Intelligence. Morgan Kaufman, San Mateo, CA, pages 271-288.

SINCLAIR, J., editor. 1987. Looking Up: An Account of the COBUILD Project in Lexical Computing. Collins, London.

SKOROCHOD'KO, E. F. 1972. Adaptativemethods of automatic abstracting and indexing. In C. V. Freiman, editor, Information Processing 71: Proceedings of the IFIP Congress 71, pages 1179-1182, NorthHolland Publishing Company.

SLATOR, B.M. 1992. Sense and preference.Computer and Mathematics with Applications, 23(6/9):391-402.

SLATOR, B.M. AND WILKS Y.A. 1987.Towards semantic structures from dictionary entries. In Proceedings of the 2ndAnnual Rocky Mountain Conference on Artificial Intelligence, pages 85-96, Boulder,CO.

SLAUGHTER, M.M. 1982. Universal Languages and Scientific Taxonomy in the Seventeenth Century. Cambridge: Cambridge University Press.

SMALL, S.L. 1980. Word Expert Parsing: A Theory of Distributed Word-based Natural Language Understanding. Ph.D. thesis, Department of Computer Science, University of Maryland, September. Available as Technical Report 954.

SMALL, S.L. 1983. Parsing as cooperative distributed inference. In Margaret King, editor, Parsing Natural Language. Academic Press, London.

SMALL, S.L., Garrison W. Cottrell, and Michael K. Tanenhaus, editors. 1988. Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and ArtiJicial Intelligence. Morgan Kaufman, San Mateo, CA.

SMALL, S.L. and Charles Rieger. 1982. Parsing and comprehending with word experts (a theory and its realization). In Wendy Lenhert and Martin Ringle, editors, Strategies for Natural Language Processing. Lawrence Erlbaum and Associates, Hillsdale, NJ, pages 89-147. Sparck-Jones,

SNOW, R., JURAFSKY, D., AND NG, A. Y. 2006. Semantic taxonomy induction from heterogenous evidence. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21th Conference on Computational Linguistics (COLING-ACL, Sydney, Australia).

SNYDER, B. AND PALMER,M. 2004. The English all-words task. In Proceedings of the 3rd InternationalWorkshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain). 41–43.

SOANES, C. AND STEVENSON, A., Eds. 2003. Oxford Dictionary of English. Oxford University Press, Oxford, U.K.

SOWA, J.F. 1984. “Conceptual Structures: Information Processing in Mind and Machine”. Reading, Massachusetts: Addison-Wesley Publishing Company.

STEVENSON, M. AND WILKS, Y. 2001. The interaction of knowledge sources in word sense disambiguation. Computat. Ling. 27, 3, 321–349.

STOKOE, C., OAKES, M. J., AND TAIT, J. I. 2003. Word sense disambiguation in information retrieval revisited. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development

in Information Retrieval (Toronto, Onto., Canada). 159–166.

STRAPPARAVA, C., GLIOZZO, A., AND GIULIANO, C. 2004. Pattern abstraction and term similarity for word sense disambiguation: Irst at Senseval-3. In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain). 229–234.

SUSSNA, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In Proceedings of the 2nd International Conference on Information and Knowledge Base Management (Washington D.C.). 67–74.

SUTCLIFFE, R.F.E., A. McElligott,D. O'Sullivan, A. A. Polikarpov, L. A. Kuzmin, G. O'Neill, and J. Véronis.1996. An Interactive Approach to the Creation of a Multilingual ConceptOntology for Language Engineering. In Proceedings of the Workshop "Multilingualityin the Software Industry," European Conference on Arti~'cial Intelligence, ECAI'96,Budapest University of Economics, Budapest, Hungary, August.

SUTCLIFFE, R.F.E., D. O'Sullivan, A. A. Polikarpov, L. A. Kuzmin,A. McElligott, and J. V~ronis. 1996. IWNR--Extending a public multilingualtaxonomy to Russian. In Proceedings of the Workshop "Multilinguality in the Lexicon,"AISB Second Tutorial and Workshop Series, pages 14-25, University of Sussex,Brighton, UK, March/April.

ten HACKEN, P. 1990. Reading distinctionin machine translation. In Proceedings of the 12th International Conference onComputational Linguistics, COLING'90, volume 2, pages 162-166, Helsinki,Finland, August.

TOWELL, G. AND VOORHEES, E. 1998. Disambiguating highly ambiguous words. Computat. Ling. 24, 1, 125–145.

TRATZ, S., SANFILIPPO, A., GREGORY, M., CHAPPELL, A., POSSE, C., AND WHITNEY, P. 2007. PNNL: A supervised maximum entropy approach to word sense disambiguation. In Proceedings of the 4th InternationalWorkshop on Semantic Evaluations (SemEval, Prague, Czech Republic). 264–267.

TSATSARONIS, G., VAZIRGIANNIS, M., AND ANDROUTSOPOULOS, I. 2007. Word sense disambiguation with spreading activation networks generated from thesauri. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI, Hyderabad, India). 1725–1730.

TUFIS, D., CRISTEA, D., AND STAMOU, S. 2004. Balkanet: Aims, methods, results and perspectives. A general overview. Romanian J. Sci. Tech. Inform. (Special Issue on Balkanet) 7, 1-2, 9–43.

TUFIS,D., ION, R., AND IDE,N. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering, and aligned WordNets. In Proceedings of the 20th International Conference on Computational Linguistics (COLING, Geneva, Switzerland).

TURING, A. M. 1950. Computing machinery and intelligence. Mind 54, 443–460.

van der Eijk, Pim. 1994. Comparativediscourse analysis of parallel texts. Second Annual Workshop on Very Large Corpora (WVLC2), pages 143-159, Kyoto, Japan, August.

VAN DONGEN, S. 2000. Graph Clustering by Flow Simulation, Ph.D. dissertation. University of Utrecht, Utrecht, The Netherlands.

VEENSTRA, J., DEN BOSCH, A. V., BUCHHOLZ, S., DAELEMANS, W., AND ZAVREL, J. 2000. Memory-based word sense disambiguation. Comput. Human. 34, 1–2.

V´ERONIS, J. 2004. Hyperlex: Lexical cartography for information retrieval. Comput. Speech Lang. 18, 3, 223–252.

V´ERONIS, J. AND IDE, N. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In Proceedings of the 13th International Conference on Computational Linguistics (COLING, Helsinki, Finland). 389–394.

VICKREY, D., BIEWALD, L., TEYSSIER, M., AND KOLLER, D. 2005. Word sense disambiguation for machine translation In Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP, Vancouver, B.C., Canada). 771–778.

VOORHEES, E. M. 1993. Using wordnet to disambiguate word senses for text retrieval. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, PA). 171–180.

VOSSEN, P., Ed. 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer, Dordrecht, The Netherlands.

WALTZ, D.L. and POLLACK J.B. 1985. Massively parallel parsing: A stronglyinteractive model of natural language interpretation. Cognitive Science, 9:51-74.

WEISS, S. 1973. “Learning to disambiguate.” Information Storage and Retrieval, 9.

WIERZBICKA, A. 1980. Lingua Menalis: The Semantic Analysis of Natural Language. New York: Academic Press.

WEAVER, W. 1949. Translation. In Machine Translation of Languages: Fourteen Essays (written in 1949, published in 1955), W. N. Locke and A. D. Booth, Eds. Technology Press of MIT, Cambridge, MA, and John Wiley & Sons, New York, NY, 15–23.

WEINREICH, U. 1980. On Semantics.University of Pennsylvania Press.

WIDDOWS, D. AND DOROW, B. 2002. A graph model for unsupervised lexical acquisition. In Proceedings of the 19th International Conference on Computational Linguistics (COLING, Taipei, Taiwan). 1–7.

WILKS, Y.A. 1968. On-line semantic analysis of English texts. Mechanical Translation, 11(3-4):59-72.

WILKS, Y.A. 1969. Getting meaning into the machine. New Society, 361:315-317.

WILKS, Y.A. 1973. An artificial intelligence approach to machine translation. In Roger Schank and Kenneth Colby, editors, Computer Models of Thought and Language. W. H. Freeman, San Francisco, pages 114-151.

WILKS, Y.A. 1975a. Primitives and words. In Proceedings of the Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing, pages 42-45, Cambridge, MA, June.

WILKS, Y. A. 1975b. Preference semantics. In E. L. Keenan III, editor, Formal Semantics of Natural Language.Cambridge University press, pages 329-348.

WILKS, Y.A. 1975c. An intelligent analyzer and understander of English. Communications of the ACM, 18(5):264-274.

WILKS, Y.A. 1975d. A preferential, pattern-seeking semantics for natural language inference. Artbqcial Intelligence, 6:53-74.

WILKS, Y.A. 1975. Preference semantics. In Formal Semantics of Natural Language, E. L. Keenan, Ed. Cambridge University Press, Cambridge, U.K., 329–348.

ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date: February 2009

WILKS, Y. AND SLATOR, B. 1989. Towards semantic structures from dictionary entries. In Proceedings of the 2nd Annual Rocky Mountain Conference on AI (Boulder, CO). 85–96.

WILKS, Y., SLATOR, B., AND GUTHRIE, L., Eds. 1996. Electric Words: Dictionaries, Computers and Meanings. MIT Press, Cambridge, MA, USA.

WILKS, Y. A., FASS, D. C., GUO, C.-M., MCDONALD, J. E., PLATE, T., AND BRIAN, B. M. 1990. Providing machine tractable dictionary tools. Mach. Transl. 5, 99–154.

WITTGENSTEIN, L. 1953. PhilosophicalInvestigations. Translated by G. E. M. Anscombe. Basil Blackwell, Oxford.

YAROWSKY, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proceedings of the 14th International Conference on Computational Linguistics (COLING, Nantes, France). 454–460.

YAROWSKY, D. 1993. One sense per collocation. In Proceedings of the ARPA Workshop on Human Language Technology (Princeton, NJ). 266–271.

Yarowsky, David. 1994a. Decision lists forlexical ambiguity resolution: Application to accent restoration in Spanish andFrench. In Proceedings of the 32nd Annual Meeting, pages 88-95, Las Cruces, NM.Association for Computational Linguistics.

Yarowsky, David. 1994b. A comparison ofcorpus-based techniques for restoring accents in Spanish and French text. InProceedings of the 2nd Annual Workshop on Very Large Text Corpora, pages 19-32, LasCruces, NM.

YAROWSKY, D. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Las Cruces, NM). 88–95.

YAROWSKY, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (Cambridge, MA). 189–196.

YAROWSKY, D. 2000. Hierarchical decision lists for word sense disambiguation. Comput. Human. 34, 1-2, 179–186.

YAROWSKY, D. AND FLORIAN, R. 2002. Evaluating sense disambiguation across diverse parameter spaces. J. Nat. Lang. Eng. 9, 4, 293–310.