Chapter 5

Building a Knowledge Base Using Generative Lexicon for Domain-Specific Compound Nouns Interpretation

5.1 Introduction

The previous chapter shows the importance of a rich linguistic resource encoded with lexical semantic knowledge to use with machine learning to improve the accuracy of the model in solving compound noun interpretation problems. This chapter proposes a knowledge-based method to solve compound noun interpretation using the generative lexicon framework. The idea of the GL theory is that semantic knowledge is encoded in the lexicon, and it provides a framework to represent the meaning of a word. The domain-specific compound nouns from the Ayurveda domain are taken for this purpose. We have developed a database of domain-specific compound nouns with their semantic knowledge representation structure. This database can be used

in further interpretation of compound nouns using statistical methods. It can also be used with machine learning algorithms to interpret the meaning of compound nouns.

Semantic processing always requires adequate knowledge representation and the ability to interpret semantically related knowledge. To understand the meaning of a concept, one needs to understand the semantic information encoded in the concept. WordNet, FrameNet, Semantic Web Network and Ontology are some of the resources developed and used to represent the semantic information of the word. Previous works on Compound noun interpretation have used semantic knowledge resources as a base to be used in machine learning algorithms.

5.2 Introduction to Generative Lexicon theory

In this section we present a detailed description of the theoretical framework generative lexicon used for making knowledge base. The basic idea of the Generative Lexicon framework is to represent words and their meanings. Generative lexicon Pustejovsky & Bouillon (1995); Pustejovsky (1998)(henceforth GL) was proposed by James Pustejovsky in 1995. It developed out of a goal to provide compositional semantics for the contextual modulations in meaning that emerge in real linguistic usage. Since it was first proposed, GL has developed to account for a broad range of phenomena involving argument alternation, polysemy, type coercion, as well as discourse phenomena and metaphor. The idea of the GL theory is that semantic knowledge is encoded in the lexicon and it provided a framework to represent the meaning of a word.

The representational structure of Generative Lexicon uses four levels of representation: Argument structure, Event structure, Qualia structure and Lexical Inheritance Structure. The Argument Structure provides information about the types and number of arguments of a verb, not only the true arguments that were discussed in

generative syntax but also some other arguments. One of them is the Default argument that is often not expressed in syntax but plays an important role in the qualia structure of a noun. These arguments are associated with the events of function and creation of the nouns. For instance, when we make a compound like a *brick house*, the noun *brick* is the element with which the house is constructed. Therefore, in the argument structure of the word *house* that is represented as a Default argument associated with the event of construction. The Event Structure analyzes the event semantics of a verb by providing the number of events associated with a verb, the sub-events within an event and event-headedness in a particular construction.

Levels of representation:

1. **LEXICAL TYPING STRUCTURE**: Gives an explicit type for a word positioned within a type system for the language.
2. **ARGUMENT STRUCTURE**:The number and nature of the arguments to a predicate.
3. **EVENT STRUCTURE**:the event type of the expression and any sub eventual structure it may have.
4. **QUALIA STRUCTURE**:a structural differentiation of the predicative force for a lexical item.

The Qualia structure is that structure where the Event and Argument structure meet. Pustejovsky (2001, 2013) analyze modifying nouns in a compound word with the qualia relations of the modified noun. Often compounds are distinguished on the basis of these qualia roles. The Qualia Structure is the most important structure for analysis of the nouns. This has four types of Qualia: Formal, Constitutive, Agentive and Telic. A Formal quale classifies a word in a hierarchical system of types by relating it to its super-type. It distinguishes the object within a large domain, e.g., shape, color, orientation, magnitude, dimensionality and position. A

Constitutive Quale is what a noun consists-of or what are the parts of the noun. Constitutive qualia represents the relation between an object and its constituents or proper parts, e.g., material, weight, parts and component elements. An Agentive Quale relates a noun to the way it is created. It represents the factors involved in the origin or "bringing about" of an object, e.g., creator, artifact, natural kind and causal chain.The Telic Quale describes the function of the noun. Telic qualia represents purpose and function of the object, i.e., purpose that an agent has in performing the object and built-in function or aim which specifies certain activities.These qualia roles are often used to form a compound.

Qualia Representation:

1. **Formal qualia** provides the category of the lexical item. It is a IS-A relation.
2. **Constitutive qualia** specifies what an item is made of and sometimes becomes crucial for the nouns and the adjectives. It is a MADE-OF relation.
3. **Telic qualia** indicates the purpose of the word. It is a PURPOSE/USED-FOR relation.
4. **Agentive qualia** explains the way an item is created. It is a CREATED-BY relation.

The representation of the Generative Lexicon type structures is given in figure 5.1. The ontology of Generative Lexicon consists of three types of words:

1. Natural Type: It consists of only Formal and Constitutive Qualia. Natural objects like air and trees belong to this type.
2. Artifactual Type: Apart from the Formal and Constitutive types, it also consists of Agentive and Telic Qualia.

   E.g. book, table etc.

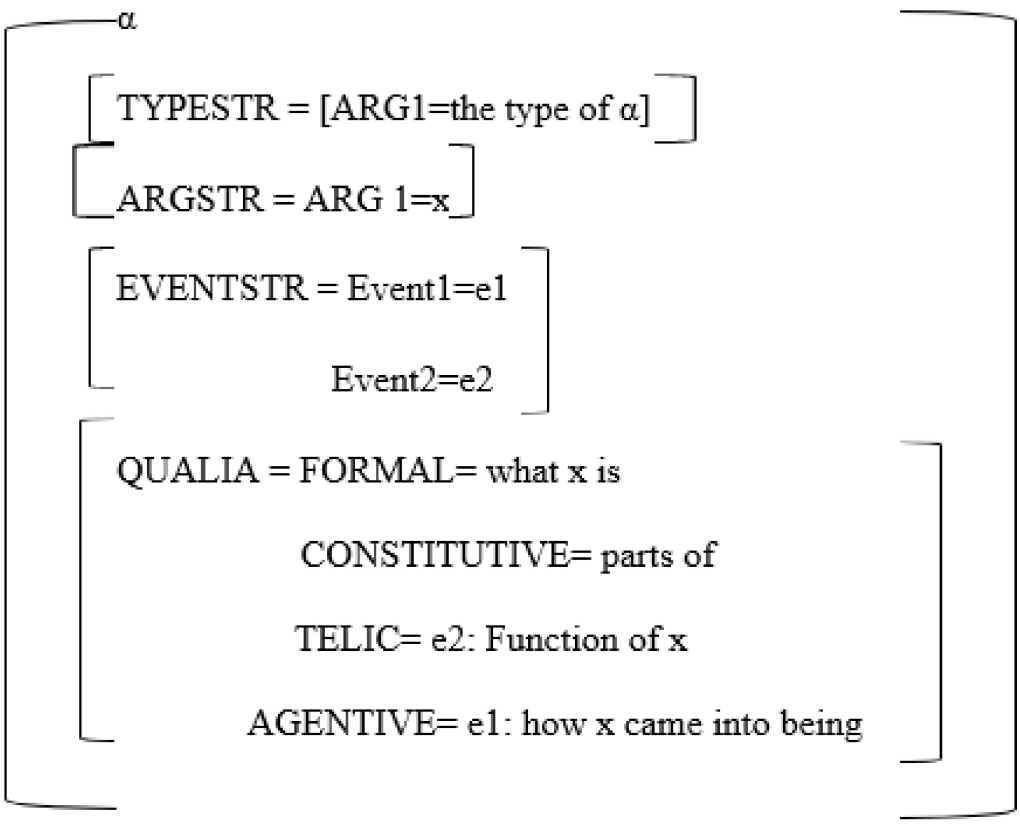

FIGURE 5.1: GL Structure of a Word

3. Complex Type: It is formed from the previous two types with a dot operation. E.g. Book: Physical object dot Information.

The operations by which a word selects another in a construction are mainly of two types.

- **Type Matching**: A type required by a function is directly satisfied by the type of the argument. E.g. The tree fell. The verb fall requires a natural type as its argument and that is satisfied by the word tree. In a compound noun, a natural object when combined with some Formal or Constitutive role of the object to make a compound, it is formed by Type matching. E.g bottle gourd and bitter gourd where bottle and bitter refer to the formal features,

shape and taste of the vegetable gourd. Similarly, an artifactual object may be combined with its Telic qualia as in reading glass 'the glass used for reading'.

- **Type Coercion**: The type required by the function is imposed on the argument it takes. This can be done in two ways.
  - **Exploitation**: Taking a part of the argument and matching with the function of the event. E.g. The car is locked where the part of the car is the door and it satisfies the function of the event lock.
  - **Qualia Introduction**: The argument is wrapped with the interpretation required by the function of the event. The type coercion changes a Natural to Artifactual Type by Qualia Introduction as in drinking water where the Telic Qualia is introduced. It can also change a count noun to mass noun as in chicken soup (soup created with chicken) where the Agentive Qualia is introduced.

5.3 Previous works with lexical knowledge bases

Compound nouns have been studied in literature using frame representation,T. W. Finin (1980); Libben & Jarema (2005). For any semantic analysis, use of semantic knowledge is unavoidable as we observe from the previous works on automatic interpretation of compound nouns Tratz & Hovy (2010); Rosario & Hearst (2001); Ponkiya et al. (2021); Fares (2019); Mezghanni & Gargouri (2017). Tratz & Hovy (2010) have used three linguistic resources, WordNet based Noun Features, Web1T corpus and Roget Thesaurus. Rosario has used a lexical ontology (MeSH tree structure) based on the medical domain text. Mezghanni & Gargouri (2017) used an ontology to extract Arabic compound nouns.Fares (2019)has used NomBank, PCEDT features, Ponkiya et al. (2021) used Frame Net based relations and Rallapalli & Paul (2012) has used PurposeNet ontology for semantic

interpretation of Compound nouns. SIMPLE-OWL(Semantic Information for Multipurpose Plurilingual Lexicons -Ontology Web Language) is an ontology based on generative lexicon framework developed by Toral et al. (2007) to be used in NLP tasks and Semantic Web. The ontology was developed from existing computational linguistics resources and converted into W3C standard ontology language.Toral et al. (2008) stated that this ontology was semantically and linguistically rich for the use in automatic semantic text processing. In Generative Lexicon Framework Qualia representation is primarily associated with the Nouns and Adjectives.Lenci et al. (2000) proposed an extended version of the Qualia Structure to include in SIMPLE-OWL for representing Nouns. In Euro WordNet, Elkateb et al. (2006) also qualia structure was proposed as an organizing principle for the top ontology. From these we can conclude that including a lexical and semantic knowledge base is a significant step to the performance of any automatic semantic interpretation task.

In the next section, we will review the existing works which have used Generative Lexicon to resolve the problem of compound noun interpretation. The earlier work done in this area by Pustejovsky & Boguraev (1993) describes a theory of lexical semantics making use of a knowledge representation framework that offers a richer, more expressive vocabulary for lexical information. Different perspectives on their representation and processing are found in Libben et al. (2003). One of the earliest works in the compound noun interpretation using GL is of Johnston & Busa (1996) focusing on the divergence of compound noun constructions of English and Italian used GL for their analysis. Other works include Bassac & Bouillon (2001); Bouillon et al. (1992) analyzed the Italian and French complex nominals using qualia structure. Kontos et al. (2000) proposed a model for developing a generative lexicon based on Greek Medical terms based on the sense of their component words. The lexicon is derived from an existing lexicon containing the terms and the definition. The ARISTA Kontos, Malagardi, et al. (2002); Kontos, Elmaoglou, & Malagardi (2002); Kontos et al. (2005) which stands for Automatic Representation Independent Syllogistic Text Analysis method is used

for developing the lexicon. Song & Zhao (2013) who have analyzed metaphorical compound nouns found in Mandarin Chinese. They proposed the compound and the metaphorical components of the compound have the same quale roles. The study found only three quale roles; FORMAL, CONSTITUTIVE and TELIC, associated with the metaphorical compound nouns. No instances of AGENTIVE quale was found. K.-S. Lee et al. (2002) also worked on cross lingual studies on the compound nouns using qualia representation. Zuoyan & Qingqing (2013) took some Chinese compounds with verbal elements and analyzed them with qualia relations. Wang et al. (2011) worked on adjectival modification to nouns in Chinese in the GL framework. Thay have used head nouns denoting events while modifiers can be or not an event. The study found that morphologically, the head and modifier is free or bound morpheme, syntactically the modifier is nominal, adjectival, verbal or numeral morpheme whereas head is always nominal morpheme. They have used the approach that the modifier is the quale role of the head and the semantic information comes from the modifier and head relations. Kriengket et al. (2007) worked on Thai compound nouns. He used those Thai compounds; the head noun of that compound refers to the scientific instruments. The representation of the compound nouns denotes the internal relations, and the properties of predicate and argument structure of the nouns. Kriengket et al. (2007) proposed that the boundary of the compounds can be calculated by analyzing the number of arguments occurring with a predicate. The result of the analysis is that the function of the head nouns mostly depicts [INSTRUMENT], [SHAPE], or [CONTAINER] relation and the argument of the nouns have two functions; The first one expresses [PURPOSE], [CHARACTERISTIC], [STATE], [METHOD], or [PROCESS]. The other expresses [PURPOSE] or [CHARACTERISTIC].

The previous work on compound noun analysis using Generative Lexicon has used a theoretical framework to understand the compound noun structure and how the compound noun is represented in the generative lexicon framework from taking the examples of compound nouns from different languages viz; French, English, Thai,

Chinese, Arabic and metaphorical compound nouns to compositional compound nouns. No work has taken the automatic interpretation of compound nouns. We are proposing an initial attempt towards the automatic interpretation of compound nouns using qualia representation of nouns based on generative lexicon framework.

5.4 Methodology of lexical knowledge base creation

This section talks about the creation and annotation of the data set in the Aayurveda domain and the creation of a knowledge base from that. Ayurveda is an ancient Indian medicine system consisting of natural treatment and healing. It uses natural medicines for the prevention and cure of diseases as well as maintaining a healthy life. The Ayurveda consists of two words: aayuh 'life' and veda 'knowledge'. Therefore, the translation of the word is 'knowledge of life'. We selected this domain for two reasons: firstly, it extends to our first dataset from the health domain. Secondly, this alternative medicine system is becoming extremely popular in India and around the world in recent times as a way to keep one fit and healthy. This domain has many compound words, as in other health and medicine domains. Many non-technical articles related to Ayurveda are found in newspapers, blogs, magazines etc. These articles, however, use some common technical terms and compound words of Ayurveda. The terms are also found in the pamphlets distributed by different ayurvedic drug companies along with their products. These terms are often used as a headword in a Compound word formation. Analysis of their Qualia Structure in terms of their constituents, purpose and creation will help to understand their meaning and predict the meanings of the compound words that are formed using them. The currently available lexicon of Ayurveda provides only a definitional description of a term that fails to relate it to other terms with which it may be associated. For example, *chuurNa, rasa, rasaayana* are some common terms of *Ayurveda* that are

used widely to make compound nouns. The head of the compounds determines with which type of noun it can be combined when making a compound noun. However, no computational work is available in this domain for lexical knowledge representation, extraction or interpretation. We thought this could be a baby step to start work in this extremely important domain.

5.4.1 Creation of Corpus

First of all we discuss the process of obtaining the corpus and making the data set from that. Ayurveda dataset creation was a very difficult and time consuming task as there was no corpus of Ayurveda available in public domain. The corpus for this work is collected from the papers of the category 'drug research' available in the Ayush portal of Government of India. We collected only the abstracts of these papers. The size of this corpus is 200000 words and the most frequent word found in the abstract corpus was *rasaayana* occurring 35 times. Other frequent terms were *chuurNa, rasa, vaTi* etc. We made a list of 30 most frequent words found in this corpus. Then making rasaayana as a keyword, we collected a corpus from the website webcorp.uk. Using bigrams, we took out all the word-sequences with rasaayana as the first or the second element. After manual checking, we found that some of these are compounds and some are not. Further, we extracted the compound words consisting of the other frequent words in the same method.

Total 400 compound words are extracted in this domain.

The most frequent ayurveda term list is provided here:

1. rasayana(chemical),
2. ghrita(liquid),
3. churNa(powder),

4. rasa(taste),
5. Ahara(food),
6. dosha(deficiency),
7. kapha,(Cough),
8. pitta(bile),
9. dravya((Drug),
10. guNa(property),
11. Aama(raw),
12. kushta(disorder),
13. pravahika(amoebiasis),
14. chikitsa(Therapy),
15. taila(oil),
16. dhatu(Tissue),
17. vati(tablets)
18. visha(poison),
19. vata(wind),
20. praNa(breath),
21. srotas(channel of circulation),
22. chakra(wheel),
23. vayu(Air),
24. agni(Fire),

25. nadi(Pulse),

26. prakriti(nature),

27. panchkarma(five procedures),

28. iccha(desire),

29. asthi(bone),

30. skandh(group/collection).

5.4.2 Annotation of Corpus

For sense annotation of these compounds, we used the available *aayurveda* dictionary[1]

. We also took help of an online document developed by a WHO project[2]

, which has listed Non-Technical Terms of aayurveda and their meanings. The link for these resources used given in the footnote at the end of the chapter with the proper numbering. We matched a compound noun with the Non-Technical list and checked the definition of that compound and based on that definition we annotated the semantic relations.

Compound	Description	Semantic Relation	Qualia Relation
audbidha dravya	dravya derived from plant	N1 is source of N2	Agentive
jangama dravya	Dravya derived from animal	N1 is source of N2	Agentive

[1]https://www.carakasamhitaonline.com/index.php
[2]https://nia.nic.in/pdf/TERMINOLOGIES.pdf

ahaara dravya	Dravya through diet	N1 is part of N2	Constitutive
ausadh dravya	Dravya taken through various form	N1 is part of N2	Constitutive
akashiya dravya	Dravya which has element akasha	N1 is part of N2	Constitutive
vayavya dravya	Dravya which has element vayu	N1 is part of N2	Constitutive
agnya dravya	Dravya which has element agni	N1 is part of N2	Constitutive
Shaktiroopa dravya	Substance which are in the form of energy	N1 is property of N2	modifier
shaktiyukta dravya	substance which possess and provide energy	N1 is property of N2	modifier
shaktihiina dravya	substance which are devoid of energy	N1 is property of N2	modifier
apya dravya	Dravya which has element prithvi	N1 is part of N2	Constitutive
sanshamna dravya	Used for pacifying biological humours	N2 is for N1	Telic

sanshodhana dravya	Used for removing humours from body	N2 is for N1	Telic
jvaraghana dravya	Used in treatment of fever	N2 is for N1	Telic
krimighana dravya	Used for alleviating worm infection	N2 is for N1	Telic
rasa dhaatu	Produced from the digestion of food and liquid	N is type of N2 (location)	Formal
rakta dhaatu	blood tissue	N is type of N2 (location)	Formal
maamsa dhaatu	The tissue that covers all organs and is related to strength and stability.	N is type of N2 (location)	Formal
meda dhaatu	fatty tissues of the body	N is type of N2 (location)	Formal
Asthi dhaatu	bone tissue	N is type of N2 (location)	Formal

Majjaa dhaatu	Majja Dhatu refers to the nervous system	N is type of N2 (location)	Formal
Shukra dhaatu	male and female reproductive tissues	N is type of N2 (location)	Formal
madhura rasa	sweet taste	N1 is property of N2	Formal
amla rasa	sour taste	N1 is property of N2	Formal
lavana rasa	salty taste	N1 is property of N2	Formal
kaTu rasa Hot and spicy taste	N1 is property of N2	Formal	
tikta rasa	bitter taste	N1 is property of N2	Formal
Kashaaya rasa(kasailaa)	astringent taste	N1 is property of N2	Formal

TABLE 5.1: Frequent Compound Nouns found in Ayurveda Corpus

For all the extracted 400 Ayurveda compounds using the aforementioned method

we annotated them with two semantic relations. First method was based on their descriptions between the constituents using our predefined semantic relation set developed for a domain specific of 20 relations. Most of the concepts are related to *chuurNa, vaTi, ghrita, rasa* and have a proper name. The name of the compound is formed using the name of constituents from which the *chuurNa* is made of. Therefore, N1 mostly contains names of herbs, fruits, i.e. proper names, hence NAME relation. The second most frequent relation was the PURPOSE relation. The first constituent is the thing which is used for the second constituent and that is why it is defined as Purpose relation. The third relation was TYPE-of relation.

Then we used qualia relation and annotated the relation between the constituents based on the qualia roles of modifier- head structure. We found that the most frequent quale roles are FORMAL (33%), TELIC (26%) and CONSTITUTIVE (29%). A small amount of AGENTIVE quale (around 13%) was present in our corpus. The graph in the figure 5.2 illustrates the frequency distribution of qualia relations in our corpus.

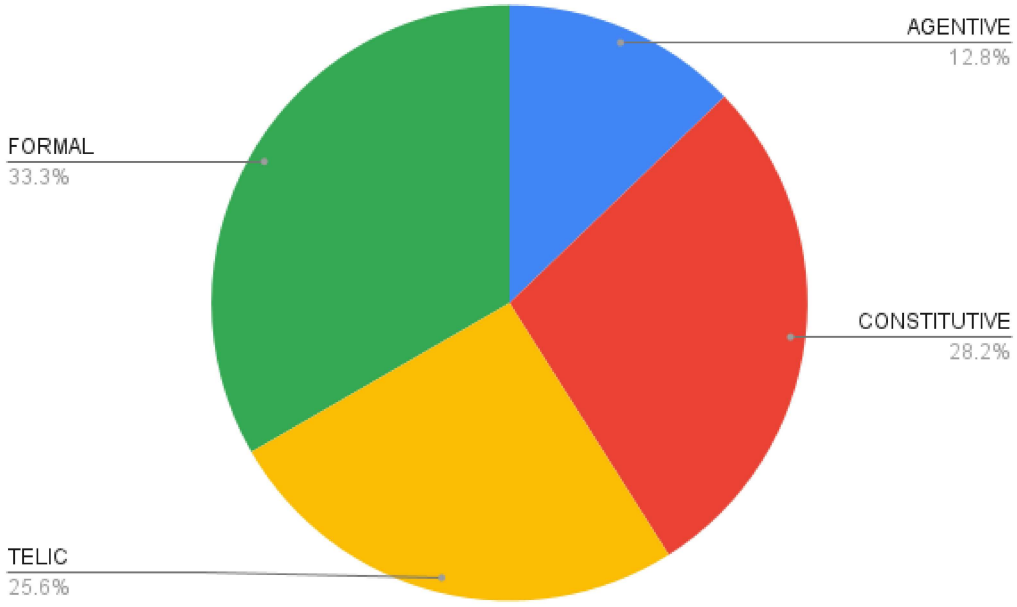

FIGURE 5.2: Frequency Distribution of Qualia relation in Compound noun data

The Table 5.1 represents some compound nouns and their semantic relation with qualia relations.

5.5 Generative Lexicon and Semantics of Compound nouns

Generative Lexicon is a lexical semantic theory which works on the concept of how we are able to give an infinite number of senses to words with finite means. It is used to represent the polysemous nature of words in the lexicon. The lexicon is an active and central component in the linguistic description. According to Generative Lexicon theory, lexical meaning is fundamentally decomposition. It is represented using the level of representation in the generative lexicon framework. Compound nouns having head modifier structures are mostly compositional. Therefore, the generative lexicon framework is used to represent the compound nouns and is useful in understanding the meaning of the compound nouns. If we take an example of two compound nouns in Hindi *naukaa vihaara* and *bauddha vihaara*, the meaning of naukaa vihaara (sailing in boat) is boating in English and bauddha vihaara (a place where monks live), i.e., abbey in English. The word vihaara is polysemous, the first vihaara means trip and the second sense is a place where the monks live. The generative lexicon accounts this word as two words and thus shows two different representations. One sense of vihaara is the process which is formed from the verb *vihaara karnaa* (to travel) and another is an artifact or a building. The process takes an argument which is used for the process to happen, therefore, it will take naukaa as an instrument argument by which the process of vihaara is being conducted. While the other sense of vihaara is a type of house and the house is used to live in. So it takes an argument like persons who live in the place or the Telic quale is exploited. Vihaara as a building is built for the buddhist monks.

In this section, we present a case study with the word rasaayana on how a GL representation using qualia structure may be used for the interpretation of the compound nouns. The term rasaayaana literally means 'augmentation of rasa'. This is a process of total rejuvenation of the body and mind done by different ways such as following certain food habits, changing lifestyle or administering some medicine. The word is also used as a product obtained at the end of this process. Therefore, rasaayana is used as a dot or complex type Process Result. In some of the compounds, the Process is exploited and in some others, the Result.

First of all, we annotated the nominal compounds with the word rasaayana with the semantic relation set developed by us and described in chapter 3. After that, we provided the Qualia structure of the head noun rasaayana in the compound words. We also observed some patterns of compound formation with this word and constructed some rules based on these patterns for the identification of the relations.

5.5.1 Qualia Representation and description of *rasaayana*

The compounds as described in the texts can be classified according to the Qualia structure of the words. We found that the noun with which this word is composed to make a compound are either its constituent part or its purpose or the way through which it is created. The table in the figures 5.3 & 5.4 illustrates some common rasaayana compound words extracted from the corpus and their meaning, relation and qualia roles.

Apart from these compounds, numerous other compounds of similar relations are often found in the texts of aayurveda with rasaayana.

Compound Noun	Gloss	Meaning	Relation	Qualia Role	Parts of Speech of N1
aacaara rasaayana	*conduct rejuve-nation*	Practising *rasaayana* through some disciplines	Modifier	N2 is created by some N1 (Agentive)	Process Noun
aahaara rasaayana	*Food rejuvenation*	Practicing *rasaayana* through some food	Modifier	N2 is created by some N1 (Agentive)	Process Noun
aajasrika rasaayana	*perpetual rejuvenation*	Daily routine based *rasaayana*	Modifier	N2 is created by some N1 (Agentive)	Adjective
naimittika rasaayana	*occasional derived from occasion rejuvenation*	*rasayana* for some specific reason (nimitta)	Purpose	N2 is for N1 (Telic)	Adjective
kaamya rasaayana	*desirable derive from desire rejuvenation*	*rasaayana* for fulfilling a desire (*kaama*)	Purpose	N2 is for N1 (Telic)	Adjective
Medhya rasaayana	*intelectual derived from intellect rejuvenation*	*rasaayana* made for increasing intellect (medhaa)	Purpose	N2 is for N1(Telic)	Adjective
kuTiipraveshika rasaayana	*hut entering rejuvenation*	*rasaayana* administered through entering a house (indoor)	Modifier	N2 is created/administered through N1(Agentive)	Adjective (from process noun)
droNipraveshika rasaayana	*a special type of wooden boat entering rejuvenation*	*rasaayana* administered through entering a special type of boat or *droNi*	Modifier	N2 is administered through N1 (Agentive)	Adjective (from process noun)
vaataatapika rasaayana	*(in)air and sunlight rejuvenation*	*rasaayana* administered in the outside in air and sunlight	Modifier	N2 is created/administered Through N1(Agentive)	Adjective (from common noun)

FIGURE 5.3: Relations and Qualia roles in the compound words with 'rasaayana' as head

kharaliya rasaayana	*made from mortar rejuvenation*	rasaayana created in *kharala* (mortar and pestle)	Modifier	N2 is created with N1 (Agentive)	Adjective (from common artifact oun)
parpati rasaayana	*thin flake rejuvenation*	Thin flake like *rasaayana*	Formal	N2 is like N1 (Formal)	Common Noun (artifact)
pottali rasaayana	*cloth packet rejuvenation*	*rasaayana* prepared in a cloth	Modifier	N2 is created in N1 (Agentive)	Common noun (artifact)
kuupipakva rasaayana	*glass bottle heating rejuvenation*	*rasaayana* prepared in a bottle	Modifier	N2 is created with N1 (Agentive)	Adjective (from process noun)
aamlakii rasaayana	*Indian gooseberry rejuvenation*	*rasaayana* made of *aamlakii*	Constitutive	N2 is made of N1(Constitutive)	(food) Noun
triphalaa rasaayana	*three fruits rejuvenation*	*rasaayana* made of *triphalaa*	Constitutive	N2 is made of N1(Constitutive)	(food) Noun
vacha rasaayana	*sweet flag(Acorus calamus) rejuvenation*	*rasaayana* made of *vacha*	Constitutive	N2 is made of N1 (Constitutive)	(food) noun
guduchi rasayaana	*A type of herb(heart leaved moonseed) rejuve-nation*	*rasaayana* made of *guduchi*	Constitutive	N2 is made of N1 (Constitutive)	(food) noun
cikitsaa rasaayana	*Treatment rejuvenation*	*rasaayana* for treatment	Purpose	N2 is for N1 (Telic)	process noun

FIGURE 5.4: Relations and Qualia roles in the compound words with 'rasaayana' as head

5.5.2 Qualia Structure of the word *rasaayana* and some compounds with it

In this section we have provided the qualia structure representation of the term rasaayana. When the word rassayana selects an adjective made from a process noun

(kuuTipraveshika, droNipraveshika etc) or an adjective from the common noun of artifact type (kharaliya) as the modifier, the role is Agentive. If it is a process noun like aahaara 'administration of food' and aacaara 'administration of some disciplines' or a common noun of artifact type (poTTali 'a cloth') through which the N2 is created, the qualia role is also agentive. The GL representation of the word rasaayana is given in the figure 5.5.

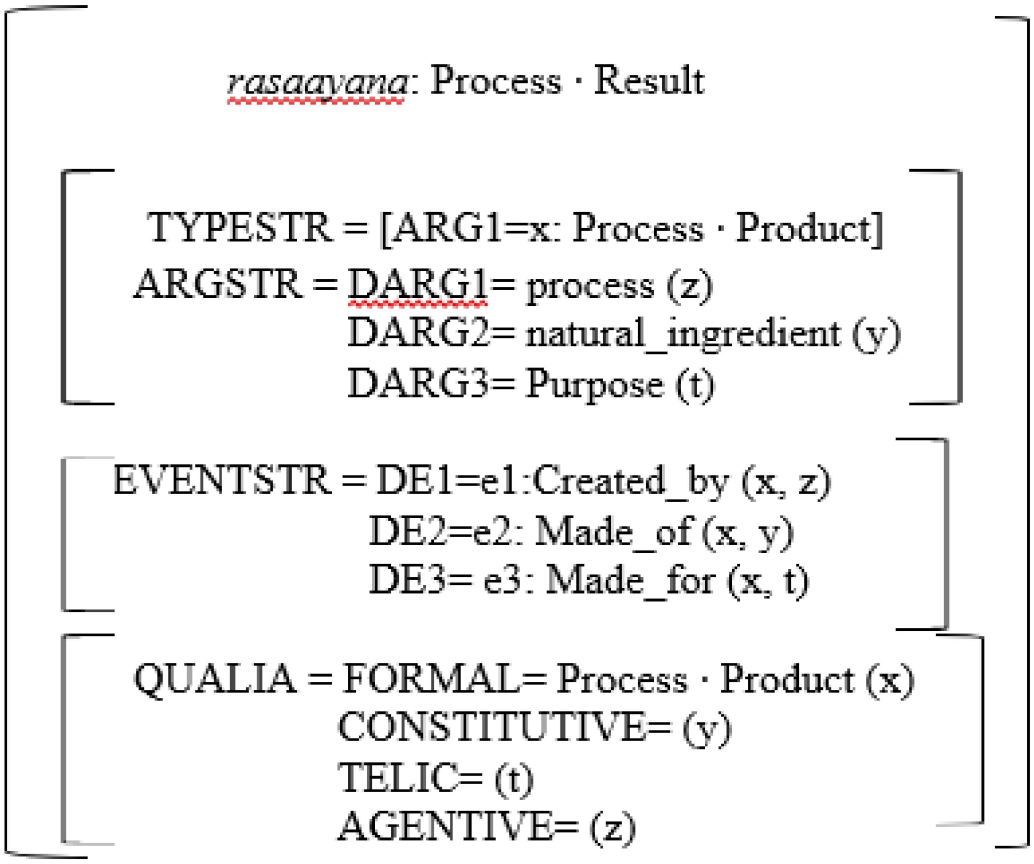

FIGURE 5.5: GL representation of the word *rasaayana*

When N2 is created by the process of N1, the resulting compound exploits the Agentive quale of N2. The syntactic category of the N1 is either a process noun or an adjective from the noun. The resulting compound is also a process noun. The illustrative examples follow.

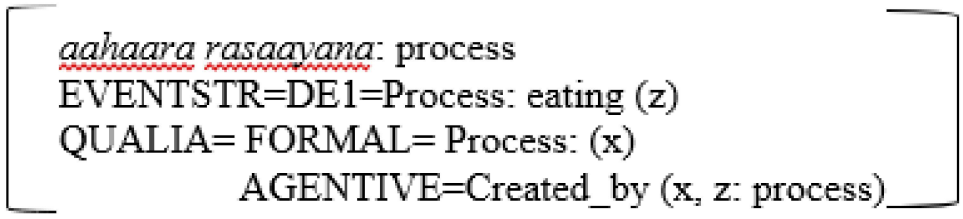

FIGURE 5.6: GL structure of the word *aahaara rasaayana* or rasaayana administered through food

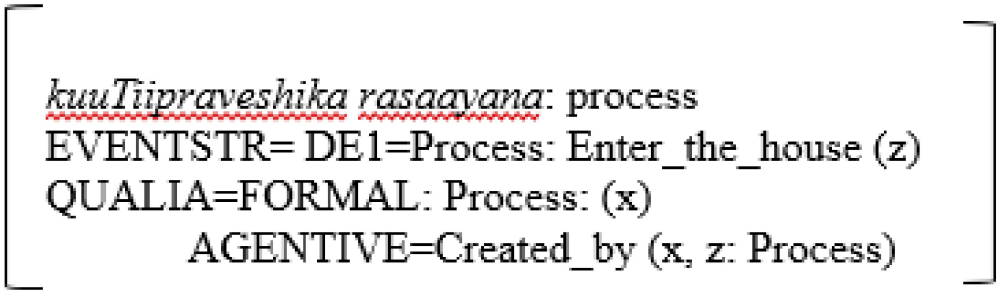

FIGURE 5.7: GL structure of the word *kuuTipraveshika rasaayana* or rasaayanaa administered inside a house

When N2 is created with the help of N1 which is a common noun of artifact type or an adjective derived from it, then the resulting compound is a product and it also exploits the agentive quale of N2. The relation between the N1 and N2 is a modifier-modified relation.

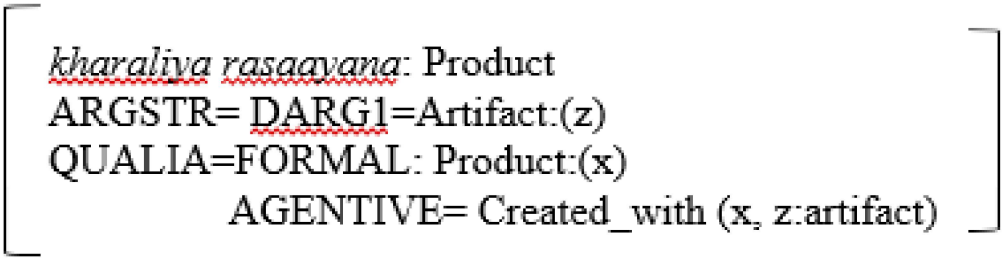

FIGURE 5.8: GL structure of the word *kharaliyarasaayana* 'the rasaayana product made with kharal or mortar and pestle

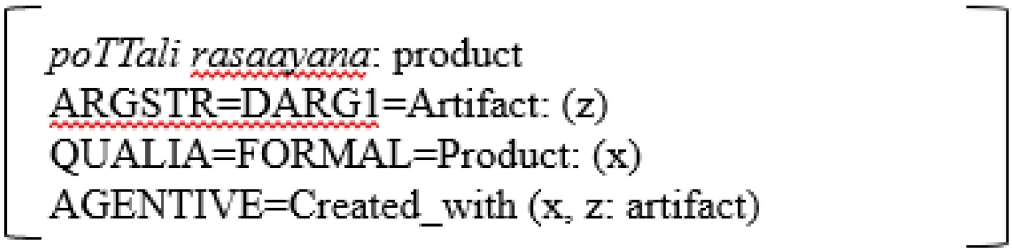

FIGURE 5.9: GL structure of the word *poTTalirasaayana* 'the rasaayana product made in poTTali or cloth'

When N2 is for (augmenting) N1, the relation between them is Purpose and the Telic Quale of N2 is exploited. The N1 is either an adjective from an abstract noun or a process noun. The examples follow.

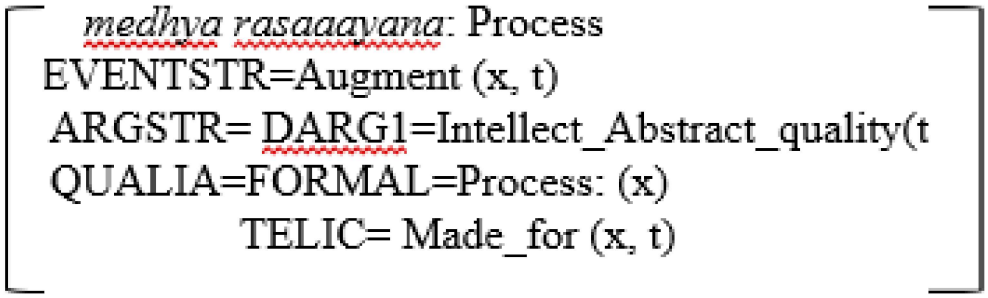

FIGURE 5.10: GL structure of the word *medhyarasaayana* ' rasaayana made for increasing intellect'

When N2 is made of N1 and N1 is a proper noun (name of a fruit or herb), the relation between the nouns is Constitutive. The resulting compound is a product.

5.5.3 Rules for identification of the relation in rasaayana compounds

We observed that the combination of the noun –noun or Adjective-Noun compounds depends upon the relations between the nouns, the argument and event structure

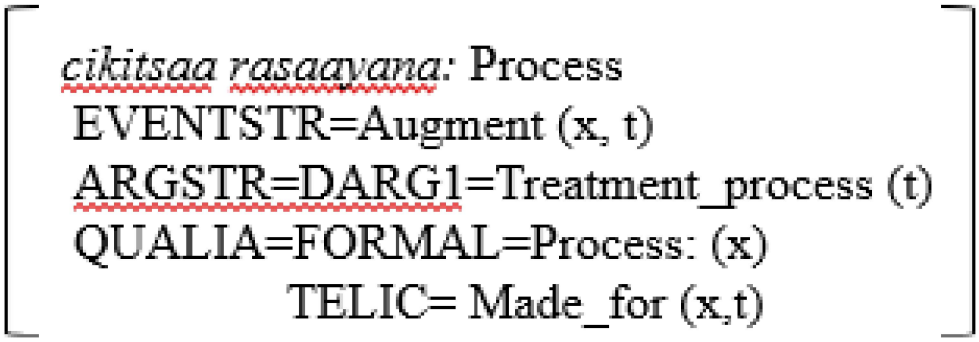

FIGURE 5.11: GL structure of the word *cikitsaarasaayana* 'rasaayana for specific treatment'

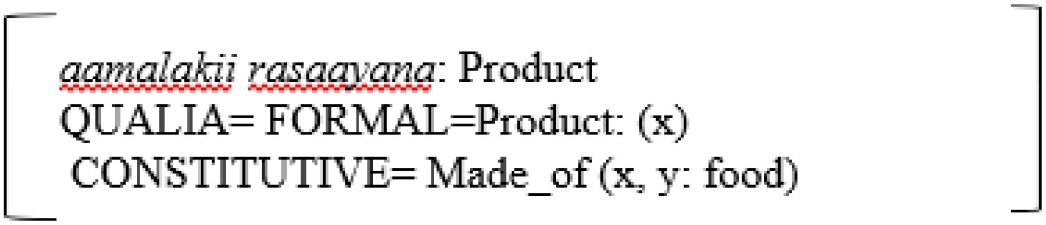

FIGURE 5.12: GL structure of the word *aamlakiirasaayana* or rasaayana made of Indian gooseberry'

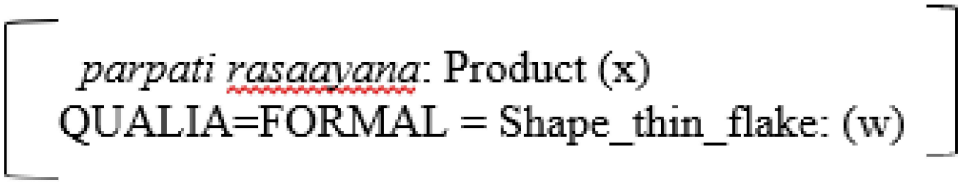

FIGURE 5.13: GL structure of the word *parpatirasaayana* 'thin flake like medicine rasaayana'

constituents of the head word or N2.

In this section we proposed some rules based on the patterns found in the above examples of compounds for computing generation of compounds with the head word N2 rasaaayana.

- If N1 of the compound is a Process noun or a common Noun or an adjective

derived from that noun , the agentive quale of N2 is exploited. The relation between them is Modifier. e.g. *aacaara rasaayana 'conduct rasaayana, poT-Tali rasayana 'cloth-packet rasaayana', kharaliya rasayan 'derived from mortal rasaayana', kuuTiipraveshika rasaayna 'rasaayana administered entering in a hut'*

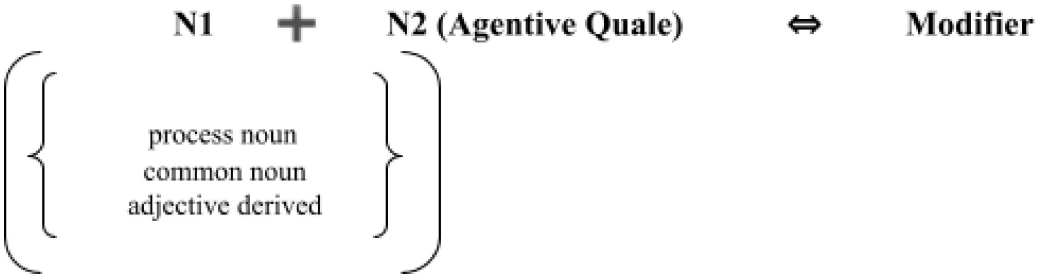

FIGURE 5.14: Rule for representing Modifier Relation

In the above examples, aacaara is a process noun, poTTali is a common noun meaning 'packet made of cloth', kharaliiya is an adjective derived from the artifact noun kharalii 'mortal' and kuuTiipraveshikais an adjective derived from the process of kuuTiipravesha 'entering a hut' . When these words occur with the head noun rasaayana, the resulting compound has the Modifier relation.

- If N1 is an Adjective from an abstract noun or a process noun exploiting the Telic Quale of N2, the relation between N1 and N2 is a Purpose relation. e.g. *medhya rasayana 'rasaayana' for (increasing) intellect'*

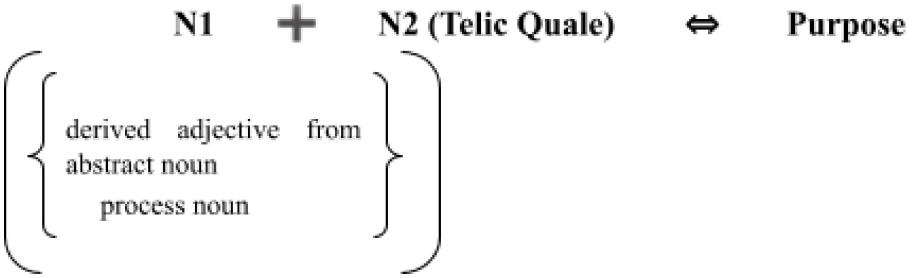

FIGURE 5.15: Rule for representing Purpose Relation

- If the N1 is a proper noun (name of fruits/herbs), the Constitutive Quale of N2 is exploited and the relation between N1 and N2 is a made-of/Constitutive. E.g.*amalaki rasayana 'Indian gooseberry rasaayana' guduchirasaayana 'heart-leaved moonseed (a type of herb, Tinospora Cordifolia) rasaayana'*

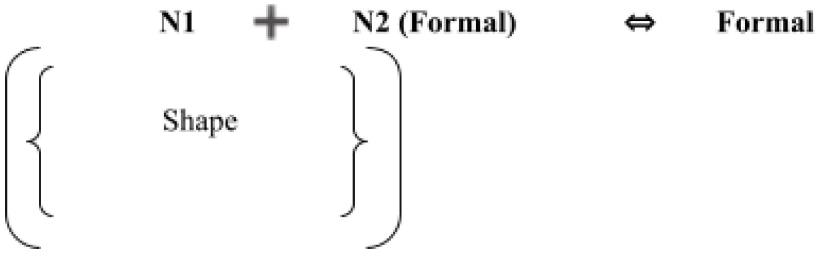

FIGURE 5.16: Rule for representing Formal Relation

- If N1 is a formal feature of N2, the resulting compound is a product type and the relation between the constituents is Formal. E.g.*parpatirasaayana 'thin flake shaped (medicine) rasaayana'*

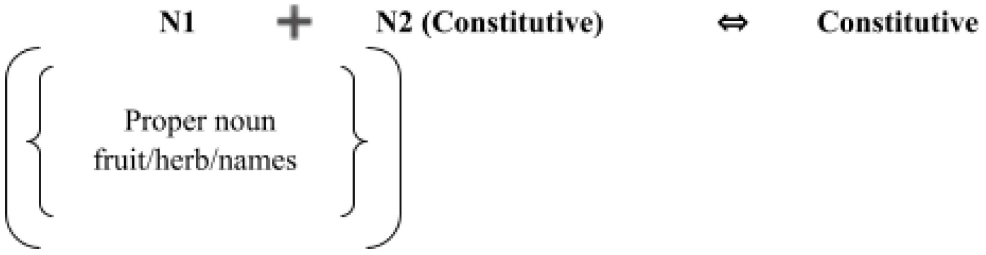

FIGURE 5.17: Rule for representing Constitutive Relation

5.6 Interpretation and Discussion

We have discussed the making of qualia structure of the terms of *aayurveda* giving an example of the term *rasaayana*. We have shown that four types of qualia roles Agentive, Telic, Constitutive and Formal exist in the x-rasaayana compounds between the constituents of the compound. We also showed that using qualia structure

of the terms one can interpret the meaning of the compound as well as predict the meaning of the compound based on different qualia relations. We have made qualia representation of 25 more terms found in our Ayurveda corpus ;some of which is presented in the appendix. This kind of representational model can work as a NLP tool and is very useful in many NLP tasks such as information retrieval, question answering etc.

We developed some rules to automatically extract the relationship between the constituents of compound nouns having modifier and head relation. For example, the compound noun *raama rasaayana* is of the structure x-rasaayana here x is the modifier which is raama and rasaayana as head. We know that raama rasaayana means the rasaayana which is made from raam naam 'chanting of the names of Lord Raama'. Here raam metaphorically is used as a constituent of rasaayana and similar to aamalaki rasaayana. Raama is a proper name and according to our third rule if N1 is the name of a fruit, a herb or proper name, the constitutive qualia gets exploited and the resulting relation will be Constitutive. Therefore, the meaning of raama rasaayana would be rasaayana made of raama or raama is a part of that rasaayana.

We have focused primarily on compound nouns having the modifier- head structure and head being the N2 in the N1-N2 structure. The qualia representations of the head words we made, we can predict which different types of modifier a head can take and thus how it can form a resulting compound noun. We showed how with qualia structure of the noun one can interpret the meaning of the compound as well as predict the meaning of the compound based on different qualia relations. The noun compound For e.g the qualia structure

We argue that the representation model is used in two ways for NLP problems. One way is used as a knowledge resource to extract features and relationships between the constituents of the compound nouns for machine learning algorithms. We have talked about this approach in this work using a knowledge base on qualia representations and rules to extract the semantic relations between the constituents. We

use this knowledge base as we have used WordNet for extracting noun features. Second, we use this knowledge representation knowledge base as a lexical resource and learn embeddings on this knowledge base for further use with machine learning to understand the semantic similarity and relation extraction tasks. The recent developments in NLP have shifted to incorporate ontologies with machine learning models and language models to better understand the semantic nature of words and textKulmanov et al. (2021). The present work did not deal with the second approach, we will work on it in future. The ontology based or knowledge representation based approach for automatic semantic analysis , data mining for domain specific texts have recently gained importance. The work of Rosario et al. (2002) on biomedical text compound noun analysis using MeSH hierarchical lexical ontology showed the applicability of ontologies for text analysis. The work of Robinson & Haendel (2020)Robinson presents a survey on the studies done on using knowledge representation and machine learning for natural language processing of medical texts.

5.7 Conclusion

Séaghdha (2008)in her thesis on compound noun interpretation argues that the rich knowledge resources like generative lexicon based representations suffer from immanent difficulties for constructing robust knowledge bases and also it affects the simplicity of the system as the system grows. It increases the system complexities. Some of the work which focuses on solving this problem is done by using a very large semantic networkHarrington & Clark (2007) and the automatic extraction of qualia structures for nounsCimiano & Wenderoth (2007); Yamada et al. (2007).

For semantic analysis world knowledge plays a crucial role and world knowledge cannot be understood by using only data. We need to know the relationships between

the entities, their semantic structures, and semantic knowledge. Recent developments in the area of artificial intelligence to produce intelligent systems have greatly focused on the semantic and pragmatic nature of the words. Therefore, a semantic web net is also developed. The recent works have also focused on the automatic information extraction of the semantic information from a large semantic resources viz, semantic web net, ontologies and wordnet using machine learning and neural network methods. Shwartz (2019) in her study about distributional semantics approach for compound noun interpretation claims that large language models trained with a rich semantic knowledge resource provides good accuracy and better interpretation for semantic studies.

The semantic relation set we developed has 20 semantic relations. The qualia relations are divided into only four relations. We know from our previous experiments, a small set of relations works better in a multi-class classification task. Therefore, if we use these qualia relations for a classification task, we hope to get a better accuracy in a machine learning task.

In this chapter we presented a generative lexicon based knowledge representation approach for compound noun interpretation tasks. We developed a knowledge base of qualia representation of compound nouns for domain specific compounds of Ayurveda domain. We proposed a theory that The modifier of the compound noun acts as the quale role of the head in a compound and qualia representation of the head noun is used to understand the compositional meaning of compound nouns.