Chapter 3

Hindi Compound Noun Semantics: Creation of Dataset and Annotation Schema

3.1 Introduction

In this chapter, we present our method of creating data for Hindi compound nouns and propose an annotation schema to annotate them. We have also calculated the inter annotator agreement to test the homogeneity of the relations so that this dataset and semantic relation be a useful resource for further use in computation of Hindi Compound noun semantics.

As understood from the discussion of the second chapter, the ultimate goal of compound noun semantics is the semantic analysis of compound nouns and understanding the meaning of the compound nouns. Understanding the meaning of compound nouns entails being able to identify the different semantic relationship found between

the constituents of the compound nouns. Semantic analysis of compound nouns includes two major problems: firstly, parsing of the compound nouns having more than two constituents and secondly, the semantic relation assignment between the constituents. Compositionality also plays a major role in compound noun analysis since there are many levels of degree of compositionality. There are fully compositional compound nouns where meaning comes from compositional meaning of the constituent, and also, there are compound nouns which have no degree of compositionality between them, like idiomatic compounds. Computational linguistics studies compound noun interpretation problems from the automatic computation point of view and tries to develop tools and models to present the compound noun semantics and provide statistical, probabilistic and machine learning models for identifying and parsing the semantic relations. Earlier works in Indian languages have primarily focused on the identification and extraction of compound nouns from the corpus except the work of Kulkarni et al. (2012) which has presented semantic relations between the constituents of Hindi compound nouns based on sanskrit grammar tradition. Hindi compound nouns are studied under the samaasa system. The Sanskrit Samaasa system has classified compound nouns into six different categories based on the nature of the constituents of the compound nouns. Previous works for Indian languages have not used any computational or statistical methods for the computation of semantic relations between the constituents of the compound nouns. The thesis proposes a model for the computation of Hindi compound noun semantics using machine learning and Generative Lexicon method.

The previous chapter presents the history of the works related to compound noun semantics in theoretical and computational linguistics. The compound noun semantics work in NLP has used different relation inventories over the last 50 years for automatic interpretation of compound nouns. However, there is no compound noun dataset available publicly specifically related to Hindi Compound nouns. In this chapter we have provided a detailed description of developing a Hindi Compound noun dataset and developed a relation set for annotation of those compound nouns.

The chapter is divided into several sections. The first section after introduction gives an account of the compound noun data extraction and creation, and the next section, 3.3, gives an account of the curation of the semantic relations set. This section also provides the samaasa-based classification of Hindi Compound Nouns and why we need other semantic relation sets for compound noun interpretation. The following section, 3.4, discusses the difference between domain-specific and domain-independent compound nouns. Furthermore, in the end, we provide an inter-annotator agreement for validating our dataset of compound nouns, and then we conclude the chapter in the last section.

3.2 Creation of Hindi Compound Noun Data

This section describes how compound noun dataset for Hindi is extracted from various resources for the analysis of compound noun semantics. Earlier work on compound noun interpretation has mainly taken into consideration English data and some has used other languages viz German, French, Arabic etc. For Hindi there was no ready dataset available for compound noun interpretation problems. The corpus for compound noun interpretation is obtained from two existing resources. First, the Hindi-Urdu domain-specific (health domain) Treebank Corpus from Technology Development for Indian Languages (TDIL) website and the second is the multiword expression list developed by the IIT Bombay. We will explain in detail the process of creating our compound noun dataset.

3.2.1 Compound Noun data from TDIL

The compound noun data is extracted from the Health domain corpus (disease data) developed at Technology Development for Indian languages. TDIL data has

two types of data: Hindi-Health domain data containing disease corpus and Hindi-Tourism Corpus. We collected Hindi-Health data. The data is annotated with a dependency label using the Paninian Dependency framework using Paninian Grammatical model. The statistics of the data is as follows. Total 37000 tokens are there in a total of 15000 sentences.

Domain: Health (Diseases)

No._of_Sentences : 1.5K

No._of_Tokens : 37K

No._of_Compound Nouns: 200

An example sentence of the data is give in figure3.1:

```
</head>
<Sentence id='51'>
1       ((      NP      <fs name='NP' drel='r6-k2:NP2'>
1.1     डेंगू   N_NNP   <fs af='डेंगू,n,m,sg,3,d,0,0' name='डेंगू' posn='10'>
1.2     बुखार  N_NN    <fs af='बुखार,n,m,sg,3,o,0,0' name='बुखार' posn='20'>
1.3     का      PSP     <fs af='का,psp,m,sg,,d,,' name='का' posn='30'>
        ))
2       ((      NP      <fs name='NP2' drel='pof:VGF'>
2.1     निदान  N_NN    <fs af='निदान,n,m,sg,3,d,0,0' name='निदान' posn='40'>
        ))
3       ((      NP      <fs name='NP3' drel='jjmod:JJP'>
3.1     माइक्रोबायलोजी          N_NNP   <fs af='माइक्रोबायलोजी,n,f,sg,3,d,0,0' name='माइक्रोबायलोजी' posn='50'>
        ))
4       ((      JJP     <fs name='JJP' drel='nmod:NP4'>
4.1     संबंधी  JJ      <fs af='संबंधी,adj,any,any,,o,,' name='संबंधी' posn='60'>
        ))
5       ((      NP      <fs name='NP4' drel='k1:VGF'>
5.1     प्रयोगशाला      N_NN    <fs af='प्रयोगशाला,n,f,sg,3,d,0,0' name='प्रयोगशाला' posn='70'>
5.2     परीक्षण N_NN    <fs af='परीक्षण,n,m,sg,3,o,0,0' name='परीक्षण' posn='80'>
5.3     द्वारा  PSP     <fs af='द्वारा,psp,,,,,,' name='द्वारा' posn='90'>
        ))
6       ((      VGF     <fs name='VGF' stype='declarative' voicetype='passive'>
6.1     किया    V_VM    <fs af='कर,v,m,sg,any,,या,yA' name='किया' posn='100'>
6.2     जा      V_VAUX  <fs af='जा,v,any,any,any,,0,0' name='जा' posn='110'>
6.3     सकता   V_VAUX  <fs af='सक,v,m,sg,any,,ता,wA' name='सकता' posn='120'>
6.4     है      V_VAUX  <fs af='है,v,any,sg,3,,है,hE' name='है' posn='130'>
        ))
7       ((      BLK     <fs name='BLK' drel='rsym:VGF'>
7.1     ।       RD_PUNC <fs af='।,punc,,,,,,' name='।' posn='140'>
        ))
</Sentence>

<Sentence id='52'>
<
```

FIGURE 3.1: An example sentence from the corpus

We manually extracted the compound nouns from this corpus. We extracted all the NP tagged part of the text which has noun sequences tagged as N_NN, N_NNP. The pattern of extraction is provided in the figure. The figure illustrates some NP has postposition tagged as PSP or Adjective tagged as JJ. We excluded all other words tagged other than N_NN, N_NNP and also some compound nouns that have hyphens between them. We removed the hyphen for the final collection. Following this method, we collected a total of 200 compound nouns from the corpus. The example sentences are given in 3.2:

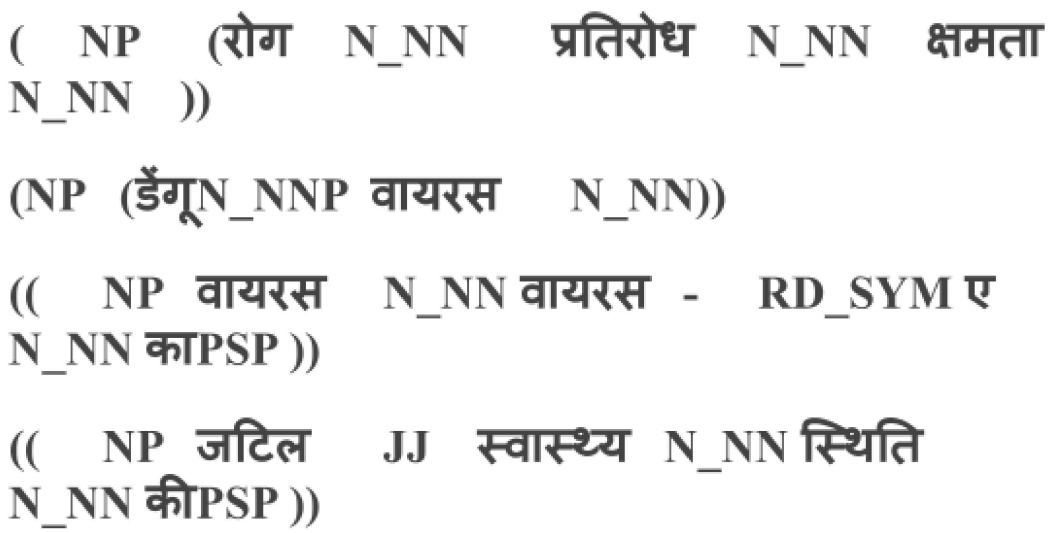

FIGURE 3.2: example sentences pattern for extracting the compound noun from the corpus

We observed 150 two noun sequence compound nouns and 50 three noun sequence compound nouns in our data. The data is presented in the appendix B.

Parsing of three noun sequences in compound nouns: In our data of health domain, we found 50 three noun sequences. We manually parsed the compound nouns based on the rules of right bracketing stated in literature. Parsing of noun sequences involves the internal structure of compound nouns. A two noun sequence compound noun has modifier-head relations between the constituents. We parsed the three noun sequences into two parts having two noun sequences in each part. Then we annotated the semantic relations between the constituents for both the parts. For example, a compound noun *jana svaasthya samasyaa* 'public health problem' is first parsed into *svaaasthya samasyaa* 'health problem' then *jana and svaaasthya samasyaa.*

jana svaaasthya samasyaa => [jana[svaaasthya samasya]] or[[jana svaaasthya]samasya]

3.2.2 Compound Noun data from Multi word expression list developed at IIT Bombay

The second compound noun data was extracted from the multiword expressions list developed at IIT Bombay. The list contained 12000 multiword expressions marked as compound nouns. The multiword list was collected from the TDIL corpus using statistical methods. The dataset marked as compound nouns had some words tagged as compounds and some not. From 12000 multiword expressions total 2500 were marked as compound nouns and remaining 9000 words were marked as not a compound. We checked all the data and found that from 2500 multiword constructions marked as compound around 1000 were not actually compounds but were only some random noun noun sequences or adjective noun sequences like *iodine manushya, behen raamkumari, baar australia* etc. And from other 9000 compound words some were compound nouns but marked as non compounds. Hence, we analyzed the whole data and checked all the compound nouns. We found 6000 legitimate

compound nouns in the entire dataset. For our study, we collected the total 1500 noun-noun sequences which are legitimate compound nouns. The 600 compound nouns sample was used to prepare the semantic relation set and annotation schema, and further 900 compound nouns were annotated based on the annotation schema after calculating the inter annotator agreement and checking the reliability of our annotation model using semantic relations. All the compounds had only two noun sequences.

3.3 Curation of semantic relations set

We discussed in previous chapters most of the earlier works on English Compound noun interpretation in NLP used a semantic relations based approach. We have used a similar approach for analysis of Hindi compound noun semantics. For Hindi, earlier work used samaas based classification and Kulkarni et al. (2012) works takes into account the semantics of Hindi and Marathi langauges using the Sanskrit Grammar tradition which has its own limitation. This section presents how samaas based classification is not appropriate of our work and presents the methodology for preparing the semantic relation set and explains the relations in detail.

3.3.1 Samaas Based Classification of Hindi Compound Noun Data

In the Sanskrit grammatical tradition, the compound words are classified according to the relative semantic prominence and the headness of the constituent word in a compound. They are primarily of two types: Endocentric where either of the constituents or both the constituents are head (tatpurusha with all its subtypes, avayaiibhaava and dvandva) and Exocentric where neither of the constituents is

the head (dvigu). If the second word is the head, the compound is called a tatpurusha compound. The tatpurusha compounds are further classified according to the vibhakti or morphological markers associated with the non-head word. In this classification also when coming to the shashthi tatpurusha or 6th case ending tatpurusha or genitive marked tatpurusha, we find that it is used for a variety of semantic relations of the compound nouns. For instance, raajaputra 'King's son' is a Family relation (father-son), kaaş haputtalii 'wooden toy' is a Constitutive relation, ta ulaka aa 'piece of rice' is a Whole-Part relation. Therefore, using only the case markers as an identifier, it is not possible to interpret the semantic relation between the words of a compound. The paraphrasing method of compound noun interpretation cannot disambiguate some extremely ambiguous case markers like Genitive as it is often used for paraphrasing in Indian languages. As explained in detail in Kulkarni et al work and its limitations for automatic interpretation of compound nouns. The work used paraphrasing and vibhakti marker found between the constituents to understand the meaning of compound nouns. In our data, we found that most dominating type of compounds are classified into Shasthi Tatpurusa (genitive), which , when anlyzed further provides different meanings for different compound nouns. We found a large number of genitive paraphrased compounds are of Theme type, i.e., the first noun is in theme relation with the second noun and the second noun is a verbal noun and others are Purpose, Contained-Container and Constitutive. For Example- *rakta sa caalan* 'blood circulation' is paraphrased as rakta kaa sa caalan where blood is the theme of the conjunct verb construction sa caalan karnaa (circulation do)'to circulate' from where the verbal noun is generated. Similar examples are *'aids transmission', iikaa pariikşa* 'vaccine testing' etc. where the second noun is a verbal noun and the first noun is the theme or object of the verb. The resulting compound noun is a process noun as the head verbal noun is also a process. Hence, we chose semantic relation based approach for understanding the semantics of Hindi Compound Nouns. This is useful in developing application based tools in NLP and Information Retrieval, Data Mining etc.

3.3.2 Procedure for Preparation of Semantic Relations

For preparing a semantic relation set for Hindi compound nouns, Rosario et al. (2002) relation set of total 48 relations and Tratz & Hovy (2010) set of total 35 semantic relations were taken into consideration. Rosario's relations are domain specific based on biomedical texts and relations of Tratz and Hovy are prepared for English general domain data.

We have two types of compound noun data: Health domain and General domain data. We used the Rosario relation set as reference for our health domain data and Tratz and Hovy relation set for our General domain data.

The relation set is given in the figure 3.3 & 3.4.

The procedure we followed for preparing our semantic relation set is as follows:

1. First we prepared the semantic relations using Health data of 200 compound nouns.
2. After that, we prepared the semantic relations for general domain data consisting of 600 compound nouns.

- We identified the meaning of compound nouns based on their context sentence and using our language and world knowledge. We checked the context in which the compound nouns are occurring and identified the possible semantic relations between the constituents of the compounds.
- The health data is not very domain specific, the data is from disease corpus, it was understandable for a native speaker of Hindi.
- We cross checked our understanding of the meaning of compound nouns and its associated semantic relations using google search and identifying the meaning

Name	N	Examples
Wrong parse (1)	109	exhibit asthma, ten drugs, measure headache
Subtype (4)	393	headaches migraine, fungus candida, hbv carrier, giant cell, mexico city, t1 tumour, ht1 receptor
Activity/Physical process (5)	59	bile delivery, virus reproduction, bile drainage, headache activity, bowel function, tb transmission
Ending/reduction	8	migraine relief, headache resolution
Beginning of activity	2	headache induction, headache onset
Change	26	papilloma growth, headache transformation, disease development, tissue reinforcement
Produces (on a genetic level) (7)	47	polyomavirus genome, actin mrna, cmv dna, protein gene
Cause (1-2) (20)	116	asthma hospitalizations, aids death, automobile accident heat shock, university fatigue, food infection
Cause (2-1)	18	flu virus, diarrhoea virus, influenza infection
Characteristic (8)	33	receptor hypersensitivity, cell immunity, drug toxicity, gene polymorphism, drug susceptibility
Physical property	9	blood pressure, artery diameter, water solubility
Defect (27)	52	hormone deficiency, csf fistulas, gene mutation
Physical Make Up	6	blood plasma, bile vomit
Person afflicted (15)	55	aids patient, bmt children, headache group, polio survivors
Demographic attributes	19	childhood migraine, infant colic, women migraineur
Person/center who treats	20	headache specialist, headache center, diseases physicians, asthma nurse, children hospital
Research on	11	asthma researchers, headache study, language research
Attribute of clinical study (18)	77	headache parameter, attack study, headache interview, biology analyses, biology laboratory, influenza epidemiology
Procedure (36)	60	tumor marker, genotype diagnosis, blood culture, brain biopsy, tissue pathology
Frequency/time of (2-1) (22)	25	headache interval, attack frequency, football season, headache phase, influenza season
Time of (1-2)	4	morning headache, hour headache, weekend migraine
Measure of (23)	54	relief rate, asthma mortality, asthma morbidity, cell population, hospital survival
Standard	5	headache criteria, society standard
Instrument (1-2) (33)	121	aciclovir therapy, chloroquine treatment, laser irradiation, aerosol treatment
Instrument (2-1)	8	vaccine antigen, biopsy needle, medicine ginseng
Instrument (1)	16	heroin use, internet use, drug utilization
Object (35)	30	bowel transplantation, kidney transplant, drug delivery
Misuse	11	drug abuse, acetaminophen overdose, ergotamine abuser
Subject	18	headache presentation, glucose metabolism, heat transfer
Purpose (14)	61	headache drugs, hiv medications, voice therapy, influenza treatment, polio vaccine
Topic (40)	38	time visualization, headache questionnaire, tobacco history, vaccination registries, health education, pharmacy database
Location (21)	145	brain artery, tract calculi, liver cell, hospital beds
Modal	14	emergency surgery, trauma method
Material (39)	28	formaldehyde vapor, aloe gel, gelatin powder, latex glove,
Bind	4	receptor ligand, carbohydrate ligand
Activator (1-2)	6	acetylcholine receptor, pain signals
Activator (2-1)	4	headache trigger, headache precipitant
Inhibitor	11	adrenoreceptor blockers, influenza prevention
Defect in Location (21 27)	157	lung abscess, artery aneurysm, brain disorder

FIGURE 3.3: Rosario Relation Set

used in the context of several sentences provided in the result of google search engine. Thus we marked most possible semantic relations.

- We observed that some of the relations provided in the Rosario and Tratz and Hovy are used as it is in our data with little modification based on our Hindi

Category Name	% Example	Approximate Mappings
Causal Group		
COMMUNICATOR OF COMMUNICATION	0.77 court order	$\supset$BGN:Agent, $\supset$L:Act$_a$+Product$_a$, $\supset$V:Subj
PERFORMER OF ACT/ACTIVITY	2.07 police abuse	$\supset$BGN:Agent, $\supset$L:Act$_a$+Product$_a$, $\supset$V:Subj
CREATOR/PROVIDER/CAUSE OF	2.55 ad revenue	$\subset$BGV:Cause(d-by), $\subset$L:Cause$_2$, $\subset$N:Effect
Purpose/Activity Group		
PERFORM/ENGAGE_IN	13.24 cooking pot	$\supset$BGV:Purpose, $\supset$L:For, $\approx$N:Purpose, $\supset$W:Activity$\sqcup$Purpose
CREATE/PROVIDE/SELL	8.94 nicotine patch	$\propto$BV:Purpose, $\subset$BG:Result, $\propto$G:Make-Produce, $\subset$GNV:Cause(s), $\propto$L:Cause$_1\sqcup$Make$_1\sqcup$For, $\subset$N:Product, $\supset$W:Activity$\sqcup$Purpose
OBTAIN/ACCESS/SEEK	1.50 shrimp boat	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\sqcup$Purpose
MODIFY/PROCESS/CHANGE	1.50 eye surgery	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\sqcup$Purpose
MITIGATE/OPPOSE/DESTROY	2.34 flak jacket	$\supset$BGV:Purpose, $\supset$L:For, $\approx$N:Detraction, $\supset$W:Activity$\sqcup$Purpose
ORGANIZE/SUPERVISE/AUTHORITY	4.82 ethics board	$\supset$BGNV:Purpose/Topic, $\supset$L:For/About$_a$, $\supset$W:Activity
PROPEL	0.16 water gun	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\sqcup$Purpose
PROTECT/CONSERVE	0.25 screen saver	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\sqcup$Purpose
TRANSPORT/TRANSFER/TRADE	1.92 freight train	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\sqcup$Purpose
TRAVERSE/VISIT	0.11 tree traversal	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\sqcup$Purpose
Ownership, Experience, Employment, Use		
POSSESSOR + OWNED/POSSESSED	2.11 family estate	$\supset$BGNVW:Possess*, $\supset$L:Have$_2$
EXPERIENCER + COGNITION/MENTAL	0.45 voter concern	$\supset$BNVW:Possess*, $\approx$G:Experiencer, $\supset$L:Have$_2$
EMPLOYER + EMPLOYEE/VOLUNTEER	2.72 team doctor	$\supset$BGNVW:Possess*, $\supset$L:For/Have$_2$, $\supset$BGN:Beneficiary
CONSUMER + CONSUMED	0.09 cat food	$\supset$BGNVW:Purpose, $\supset$L:For, $\supset$BGN:Beneficiary
USER/RECIPIENT + USED/RECEIVED	1.02 voter guide	$\supset$BNVW:Purpose, $\supset$G:Recipient, $\supset$L:For, $\supset$BGN:Beneficiary
OWNED/POSSESSED + POSSESSION	1.20 store owner	$\approx$G:Possession, $\supset$L:Have$_1$, $\approx$W:Belonging-Possessor
EXPERIENCE + EXPERIENCER	0.27 fire victim	$\approx$G:Experiencer, $\propto$L:Have$_1$
THING CONSUMED + CONSUMER	0.41 fruit fly	$\supset$W:Obj-SingleBeing
THING/MEANS USED + USER	1.96 faith healer	$\approx$BNV:Instrument, $\approx$G:Means$\sqcup$Instrument, $\approx$L:Use, $\subset$W:MotivePower-Obj
Temporal Group		
TIME [SPAN] + X	2.35 night work	$\approx$BNV:Time(At), $\supset$G:Temporal, $\propto$L:In$_a$, $\approx$W:Time-Obj
X + TIME [SPAN]	0.50 birth date	$\supset$G:Temporal, $\approx$W:Obj-Time
Location and Whole+Part/Member of		
LOCATION/GEOGRAPHIC SCOPE OF X	4.99 hillside home	$\approx$BGV:Locat(ion/ive), $\propto$L:In$_a\sqcup$From$_a$, B:Source, $\approx$N:Location(At/From), $\approx$W:Place-Obj$\sqcup$PlaceOf/Origin
WHOLE + PART/MEMBER OF	1.75 robot arm	$\supset$B:Possess*, $\approx$G:Part-Whole, $\supset$L:Have$_2$, $\approx$N:Part, $\approx$V:Whole-Part, $\approx$W:Obj-Part$\sqcup$Group-Member
Composition and Containment Group		
SUBST/MATERIAL/INGREDIENT + WHOLE	2.42 plastic bag	$\subset$BNVW:Material*, $\propto$GN:Source, $\propto$L:From$_a$, $\approx$L:Have$_1$, $\propto$L:Make$_{2b}$, $\propto$N:Content
PART/MEMBER + COLLECT/CONFIG/SERIES	1.78 truck convoy	$\approx$L:Make$_{2a,c}$, $\approx$N:Whole, $\approx$V:Part-Whole, $\approx$W:Parts-Whole
X + SPATIAL CONTAINER/LOC/BOUNDS	1.39 shoe box	$\supset$B:Content$\sqcup$Located, $\supset$L:For, $\supset$L:Have$_1$, $\approx$N:Location, $\approx$W:Obj-Place
Topic Group		
TOPIC OF COMMUNICATION/IMAGERY/INFO	8.37 travel story	$\supset$BGNV:Topic, $\supset$L:About$_{a,b}$, $\supset$W:SubjectMatter, $\subset$G:Depiction
TOPIC OF PLAN/DEAL/ARRANGE/RULES	4.11 loan terms	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
TOPIC OF OBSERVATION/STUDY/EVAL	1.71 job survey	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
TOPIC OF COGNITION/EMOTION	0.58 jazz fan	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
TOPIC OF EXPERT	0.57 policy wonk	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
TOPIC OF SITUATION	1.64 oil glut	$\supset$BGNV:Topic, $\propto$L:About$_a$
TOPIC OF EVENT/PROCESS	1.09 lava flow	$\supset$G:Theme, $\supset$V:Subj
Attribute Group		
TOPIC/THING + ATTRIB	4.13 street name	$\supset$BNV:Possess*, $\approx$G:Property, $\supset$L:Have$_2$, $\approx$W:Obj-Quality
TOPIC/THING + ATTRIB VALUE CHARAC OF	0.31 earth tone	
Attributive and Coreferential		
COREFERENTIAL	4.51 fighter plane	$\approx$BV:Equative, $\supset$G:Type$\sqcup$IS-A, $\approx$L:BE$_{Ident}$, $\approx$N:Type$\sqcup$Equality, $\approx$W:Copula
PARTIAL ATTRIBUTE TRANSFER	0.69 skeleton crew	$\approx$W:Resemblance, $\supset$G:Type
MEASURE + WHOLE	4.37 hour meeting	$\approx$G:Measure, $\subset$N:TimeThrough$\sqcup$Measure, $\approx$W:Size-Whole
Other		
HIGHLY LEXICALIZED / FIXED PAIR	0.65 pig iron	
OTHER	1.67 contact lense	

FIGURE 3.4: Tratz and Hovy Relations

data.

- We proposed three relations not found in any relation set: Collection,Reference and Dvandva (Copulative) based on our corpus of data.
- Dvandva compounds (Reduplicative : which can be called as variantions of Dvandva) are very frequent and important feature of word formation process

in Indian Languages.

- Following the above process we prepared our own semantic relation set consisting of 20 semantic relations for Hindi Compound nouns.

3.3.3 Semantic relations for Hindi compound noun annotation

The following table illustrates the semantic relations proposed by us for the compound nouns of Hindi and examples of each of the relations. We also provided how the N1 is related to N2 by a paraphrasing with prepositions. Paraphrasing is not possible only for the lexicalized and Name relation which is found in the WordNet or a dictionary. The table with The semantic relations and description is presented in the 3.1

Relation No.	Semantic Relation	Preposition between the constituents	Description of relation	Example
1	PURPOSE	for	N2 is for N1	SiSu biSeSjnya 'child specialist'
2	SOURCE	from	N1 is from N2	parmanu indhan 'nuclear energy'
3	LOCATION	in/on/at	N1is location of N2	rakta kainsar 'blood cancer'
4	INSTRUMENT	(used) for	N1 is used for N2	saikil yaatraa 'cycle tour'
5	TYPE OF	Type_of/ (hyponymy relation in Wordnet)	NC is type of N2	kampyutar gem 'computer game'

6	NAME	Proper name	N1 is a name for N2	'Sambalpur district'
7	TOPIC	related$_to$	N1 is related to N2	namak kaanun 'salt law'
8	DVANDVA	and	N1 and N2 both head nouns	kurtaa paajaamaa ''kurtapayjama'
9	MADE OF	of	N2 is made of N1	svarna patra 'golden leaf
10	POSSESSOR$_{POSSESSED}$	of	N1 is the possessor of N2	jahangir mahal'Jahangir palace'
11	CAUSE	of (se/kA)	N1 is the cause of N2 or N2 is the cause of N1	parjiibii roga 'parasitic disease'
12	COMPONENT-WHOLE	of	N1 is the part of N2	siimenTa roDa 'cement road'
13	CONTENT-CONTAINER	of	N1 is contained in N2	waTar baTal 'water bottle'
14	PART LOCATION	Part$_of$($alocation$)	N1 is part of a location N2	paschima jaawaa 'Western Java'
15	MODIFIER	(Modifier) of	N1 is modifier of N2	mahilaa dokTar woman doctor
16	THEME	of	N1 is direct theme or patient of N2	taiksii caalak 'taxi driver'
17	LEXICALIZED	NC found in Dictionary	braj bhumi 'brajabhuumi'	

18	COLLECTION	(Collection) of	NC is a collection of N1 and N2 is a collective noun	jantu səmuuha 'pack of animals'
19	MEASURE	(Measure) for	N2 is a measure word for N1	baiMk reTa 'bank rate'
20	AGENT	for	N1 is agent for N2	chaaatra aandolana, student protest

TABLE 3.1: Semantic Relation set proposed by us

The health domain data has only 200 compound nouns, which is not a good amount of data to use in machine learning experiments. We focused on general domain compound noun lists. After calculating the reliability and consistency of our semantic relation set we annotated the 900 more data for preparing a compound noun dataset for machine learning experiments. Also, we presented earlier that the health domain data does not contain the health terminology and the semantic relations for both the datasets are almost same. We finalized the 20 semantic relation set(section: semantic relation set) which we used in our further study. The graph shows the frequency distribution of compound nouns with their semantic relations 3.5. As the graph in figure 3.5 illustrates: Purpose, Location, Type of, Name, Topic ,Modifier, Theme and Lexicalized relations are around 70% of the whole compound .

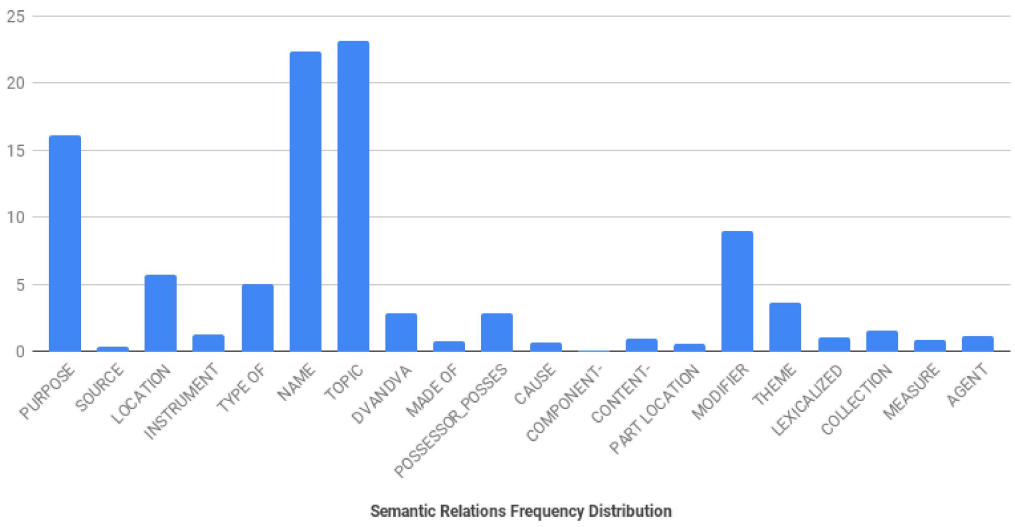

FIGURE 3.5: Semantic relations frequency distribution in our data

3.4 Semantic relations in Domain Specific vs Domain Independent Compound nouns

The present section provides an analysis of the nature of semantic relations in domain specific and domain independent compound nouns. We have observed 15 semantic relations for Health Corpus presented in appendix 1 and for the general domain, we have proposed 20 semantic relations. Most of the relations are same in both the datasets and some of them overlap. Theme semantic relation in general domain is used for the compound noun which has an activity noun N2 as the head and the modifier N1 is the theme or patient of that activity. For instance, in taxi driver taxi is the patient of the noun driver, similarly in the compound *rakta sancaaraNa* 'blood transmission', sanchaaraNa 'transmission' is an activity and the theme of the activity is rakta 'blood'. In domain specific compounds, the relation between *rakta sanchaarana* is activity relation. The earlier work of Rosario on biomedical text contains the biomedical terminology specifically used in that domain. She worked to develop a tool which is used to understand the meaning of biomedical terms and

the concepts these terms are referring to. This can be further used in data mining and information retrieval from the medical texts. Therefore, Rosario's semantic relations are domain specific and application based. Tratz and Hovy's relations are used for understanding the meaning of compound nouns using semantic relations in general. Hence, to do a fine analysis and develop an application based tool we need more fine grained classification of semantic relations particular to that domain. But our health domain data is not specialized as a medical or biomedical text, therefore, we could use the same relation set for both the domains.

3.5 Evaluating the Compound Noun Dataset

This section describes an experiment in evaluating the homogeneity of semantic relations annotation developed for Hindi Compound noun semantics. The present section shows the inter annotator agreement using the kappa coefficient for the Hindi compound noun interpretation with the experiment of labeling semantic relations found in the noun noun sequence. The agreement between the annotators using kappa Coefficient is 0.72 which is substantially good agreement. The results underline both the difficulty of the compound annotation task and the need for rigorous annotation scheme development when working with semantic data. The section is divided into four subsections. In the first subsection, we discuss the reliability checking method of inter-annotator agreement. In the next section, we have provided the data used for the evaluation and the third subsection provides the evaluation procedure. The last subsection talks about the result and discusses the agreement by comparing it with other existing literature.

3.5.1 Inter annotator agreement

Inter-annotator agreement is a measure of how well two annotators can make the same annotation decision for the category. Corpus linguistics uses inter-annotator agreement for checking the reliability of an annotation schema. Annotator agreement results illustrate how accurate and trustworthy an annotation schema is and how accurately the categories can be classified. For any classification system of machine learning, we need to train the data set with the manual classification of the compound nouns. This manual annotation finally determines the result of the classification algorithm used. Therefore, it is essential to ensure that there is a significant agreement in the classification done by the manual annotators. Inter-Annotator Agreement (IAA) method is used to check the reliability of the manual annotation for the training data. Literature uses many statistical methods to assess the inter-rater agreement between two or more annotators. There are basically two methods of calculating the inter annotator agreement ; one using percentage agreement and other one is using Cohen's kappa coefficient. Percentage agreement can be biased due to the overlapping between the categories. With the categories and data it might be possible that the majority of observations belong to some level. Therefore, Cohen's kappa is used. Kappa is predominant for calculating IAA because it takes into account the expected chance agreement which is likely to occur when annotators annotate the instances. We have developed a semantic relation set and prepared a compound noun dataset for use in automatic interpretation of Hindi compound nouns. To check the consistency of semantic relations and for preparing a reliable dataset to use in machine learning algorithms, we calculated inter-annotator agreement.

3.5.2 Data and Semantic Relation set for the inter-annotator agreement experiment

For the sample, we have taken out random 500 compound nouns having two noun sequences from a set of 1500 general domain compound noun data explained in detail in the section. The sample contains all the instances of 20 semantic relations distributed in 500 compound nouns. The sampling is done manually. We have extracted the random instances of compound nouns in proportion to the original data based on their frequency in the main dataset. For eg. in the original dataset the compound nouns which have purpose relation around 20% in our sample data it is around 17%.

The semantic relation set file contains all the semantic relations with the relation directionality of the constituents relations and the description of the relations. This compound noun data set and relation sets in different files provided to the annotators.

3.5.3 Evaluation Procedure

We have provided the data to the four selected annotators, and the complete process is described below.

Background of the annotators: All four annotators have university degrees and are native speakers of the Hindi language. Two of them have linguistics degrees, and two others have social science degrees. The annotators were instructed to annotate the relation in terms of describing the relation given in the relation set. The annotators were unknown to each other so that the judgments provided by annotators are entirely dependent on their language knowledge and annotation guidelines.

Annotators were instructed to annotate the relations using the relation set and the description of the relation set using their world knowledge about the compound and what generally that compound means in the language. Annotation is done without context information. The annotator could think of the most appropriate relation using their world knowledge and language knowledge within the given set of relations they are instructed to annotate accordingly.

The following is an example of annotation

saaikil yaatraa 'bicycle ride', ride using bicycle; Instrument relation ; **USE for**; N1 is use for N2 as instrument

Here the annotators tag the relation as an instrument as *saaikil* (bicycle) is used as an instrument for *yaatra* (travel) purpose.

3.5.4 Evaluation Method

The inter-annotator agreement was measured using Cohen's kappa. As defined by Cohen (1960), kappa was calculated using the following equation (Eq.1):

$$KappaCoefficient(k) = \frac{(p0 - pe)}{(1 - pe)} \tag{3.1}$$

Where (1-pe) gives the degree of agreement that is attainable above chance, and (p0-pe) offers the degree of the agreement achieved above chance.

We had four annotators. The best possible way to calculate the agreement was to use a permutation and combination method and then calculate the average of the score. Therefore, we calculated pairwise agreement between annotators; we made six pairs using combination formula 4C2. The first pair consists of two linguists, the second one consists of two non-linguists and the other four had one linguist and one non-linguist in each. We calculated the agreement using Cohen kappa coefficient for

each pair and then we used the average of all six kappa values to check the inter annotator agreement. The Cohen's kappa is calculated using the Cohen kappa score function in python. We calculated the agreement coefficient between the two pairs and then to get the final agreement result we calculated the average of all the kappa values.

3.5.5 Result and analysis

The kappa value obtained between two groups is given in the table 3.2

TABLE 3.2: Inter annotator agreement result

Annotator	**Annotator**	**kappa Coefficient**
Annotator 1	Annotator 2	0.974
Annotator 3	Annotator 4	0.616
Annotator 1	Annotator 3	0.816
Annotator 1	Annotator 4	0.579
Annotator 2	Annotator 3	0.795
Annotator 2	Annotator 4	0.535

The average of all the six kappa scores k = 0.719 which is equivalent to **72**.

We have got a kappa value k = 0.72, substantial agreement, which is a good level of rater agreement. The result supports the use of a relation-based approach for the interpretation of compound nouns. The disagreement between the annotators was also consistent. The result shows that the kappa value is least with the annotator 4. The most tagged relations were the purpose, modifier, type, topic, and theme. The disagreement between the annotators is found primarily in the low-frequency semantic relations. The result also shows the usefulness of a semantic relation set for the interpretation of Hindi Compound nouns. As presented in figure 3.6

The agreement result graph 3.7 demonstrates that annotator 1 and annotator 2 have got the highest kappa score and perfect agreement. The result is due to the fact that

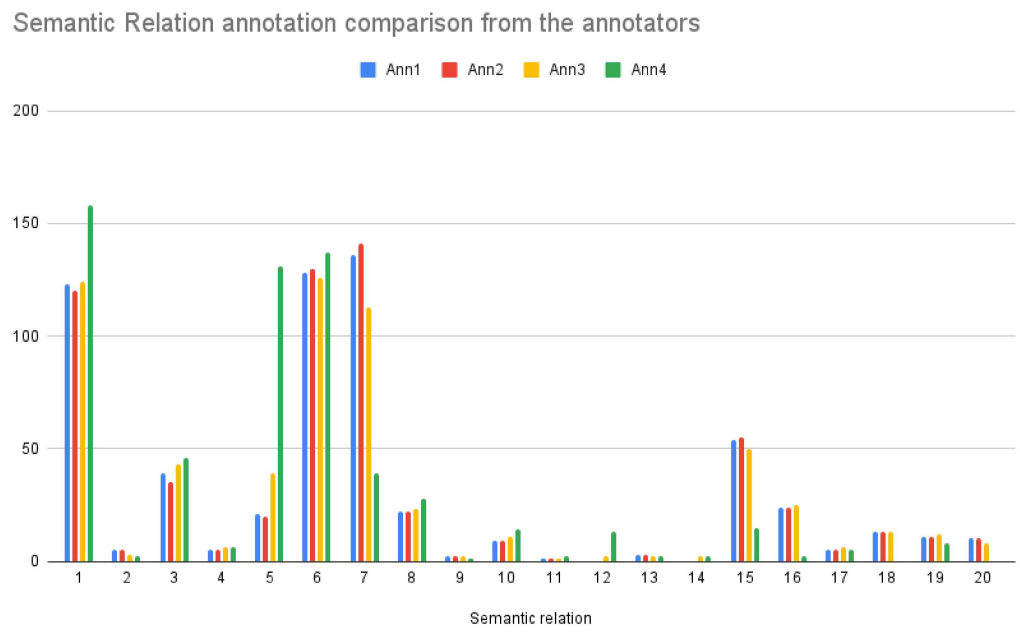

FIGURE 3.6: Semantic Relation annotation comparison from the annotators

both annotators are linguists and understand the nature of language. The lowest score is obtained with the annotator four. We argue that the one of the reasons for disagreement between the rater four and other raters is due to the individual ability. Little understanding of language or the task was not exciting as much. The semantic relation graph illustrates that the semantic relation purpose is the most annotated relation after that is the type of, topic, modifier, lexicalized etc. The agreement graph and our dataset with semantic relation frequency distribution graph shows that our semantic relations categories and dataset is consistent as in our dataset the frequent relations are the same as in annotators.

3.5.6 Prior work and Discussion

The measure of inter-annotator agreement is important, as it shows whether the annotation reliably captures the semantic information shared between users of a

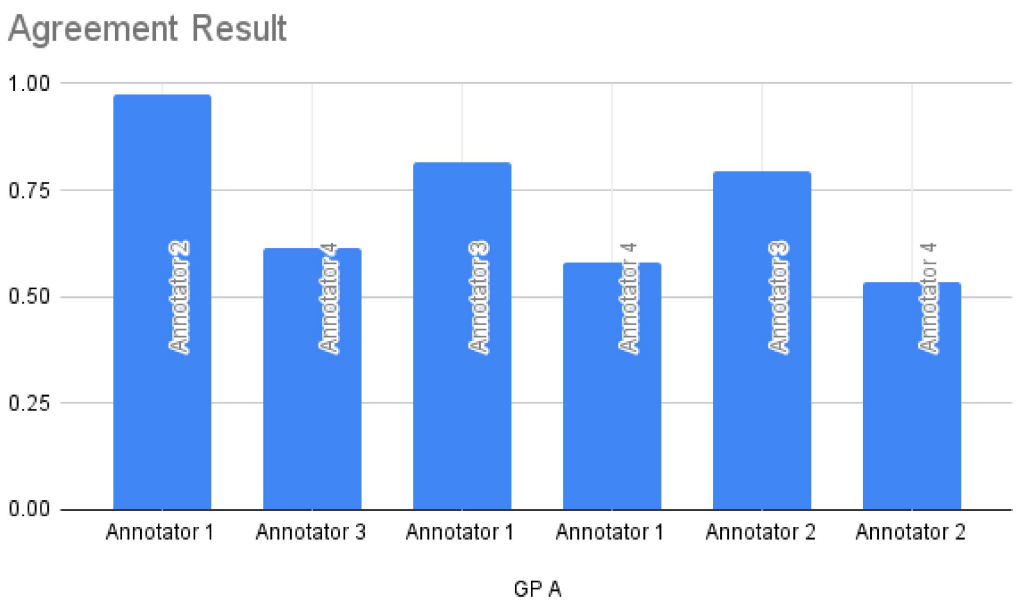

FIGURE 3.7: Comparison of Agreement results between the annotators

language. The correlation of human judgements can also calibrate our expectations of the performance that automatic methods will achieve. It would be surprising if computers performed significantly better than humans at interpreting compounds. This work appears to be the first study of calculating inter annotator for Hindi Compound noun semantics analysis.

TABLE 3.3: Comparison of Our Kappa Coefficient to other works

Dataset	**Kappa Coefficient**
Girju et al	0.58
Tratz and Hovy	0.8
Murhaf Fares	0.85
Seagadha	0.69
Our Dataset	0.72

The table illustrates the previous works on inter annotator agreement result All the results are based on English Datasets. GIRJU (2006)report agreement using LAUER

(1995) 8 prepositional labels (Kappa = 0.8) and their own 35 semantic relations (Kappa = 0.58). The highest result is of Tratz & Hovy (2010) on a general domain dataset. Fares (2019) has gotten 0.85 on his dataset which is significant as compared to others.

3.6 Conclusion

The chapter provides an analysis of Hindi compound noun semantics by giving examples of Hindi compound nouns and prepared compound noun corpus for Hindi domain specific as well as domain independent data. The chapter also discusses the semantic relations developed for Hindi and compares the domain specific and domain independent semantic relations. The chapter analyzes the difference between the semantic relations and provides the reason for these discrepancies. The chapter describes the process of inter-annotator agreement between the annotators for semantic relation sets to check the reliability of our relation set and provides a good agreement score. As for our knowledge goes this is the first ever inter-annotator agreement work done for a Hindi compound noun data set and it also compares favorably with other results in the existing literature on English data. This way, we make a Hindi compound noun standard dataset for further use in different ML experiments. The following chapter will talk about the computation of semantic relations using machine learning experiments.