

---

# Certificate

It is certified that the work contained in the thesis titled **Hindi Compound Noun Semantics using Machine Learning and Generative Lexicon** by **Vandana Dwivedi** (Roll No.: 16191501) has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy, and SOTA for the award of the degree of **Doctor of Philosophy** to the Indian Institute of Technology (Banaras Hindu University) Varanasi.

*Sanjukta* 26/12/2022

Signature of Supervisor

Dr. Sanjukta Ghosh

Associate Professor

Department of Humanistic Studies

Indian Institute of Technology

(BHU)

Varanasi - 221005, India

सहयुक्त आचार्य/Associate Professor  
मानवतावादी अध्ययन विभाग/Department of Humanistic Studies  
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology  
(काशी हिन्दू विश्वविद्यालय)/(Banaras Hindu University)  
वाराणसी-२२१००५ (उ०प्र०)/Varanasi-221005 (U.P.)

---

---

# Declaration

I, **Vandana Dwivedi**, certify that the work embodied in this thesis is my bonafide work and carried out by me under the supervision of **Dr. Sanjukta Ghosh (Department of Humanistic Studies)** from **January 2017 to December 2022**, at the **Department of Humanistic**, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.*, reported in journals, books, magazines, reports dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my work.

Date: 26/12/2022

Place: Varanasi

*Vandana Dwivedi*

Signature of Student

(Vandana Dwivedi)

## Certificate by the Supervisor

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

*Sanjukta* 26/12/2022

Signature of Supervisor

सहयुक्त आचार्य/Associate Professor  
मानवतावादी अध्ययन विभाग/Department of Humanistic Studies  
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology  
(काशी हिन्दू विश्वविद्यालय)/(Banaras Hindu University)  
वाराणसी-२२१००५ (उ०प्र०)/Varanasi-221005 (U.P.)

*Ajit Kumar Mishra*  
28/12/2022

Signature of Head of Department

(Dr. Ajit Kumar Mishra)

Head

Department of Humanistic Studies  
Indian Institute of Technology  
(Banaras Hindu University)  
VARANASI-221005 (U.P.)

---

---

# Copyright Transfer Certificate

Title of the Thesis: **Hindi Compound Noun Semantics using Machine Learning and Generative Lexicon**

Name of Student: **Vandana Dwivedi**

## Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi, all rights under copyright that may exist in and for the above thesis submitted for the award of the Doctor of Philosophy.

Date: 26/12/2022

Place: Varanasi

*Vandana Dwivedi*

Signature of Student

(Vandana Dwivedi)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

---

---

# Acknowledgments

All these PhD years have been a never-boring mixture of learning and relearning, experiments that worked in the end, experiments that didn't, stress, relief, more stress, some croquet and an awful lot of running. And at the end a thesis was produced. I want to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during this thesis work. A good number of people have contributed to getting me there, by offering advice, suggestion and feedback.

First I would like to thank My supervisor Dr. Sanjukta Ghosh, who has been an unfailing source of support, from the beginning to the end of my thesis for her continuous and immense support, guidance and encouragement. She has influenced many aspects of my writing and presentation for the better. I am grateful to my RPEC examiners, Dr. Anil Thakur and Dr. Sukomal Pal for their constructive comments, suggestions and good-humoured criticism. Their inspiration is the cause behind the successful completion of this Thesis work.

I sincerely thank Prof. R.K Mishra, Department of Electrical engineering, and Ex. Head, Department of Humanistic Studies, and Dr. A. K. Mishra , Head, Department of Humanistic Studies for providing continuous support, encouragement, and advice. I sincerely thank all the Professors, DPGC Convenors, Deans, and office staff of the Department of Humanistic Studies, India Institute of Technology (BHU) Varanasi, India. I express my gratitude to the Director, Registrars, Deans, Heads, and Student Alumni of the Indian Institute of Technology (BHU) Varanasi for this life-changing opportunity.

I would like to thanks all my friends, colleagues, friends and seniors, who constantly supported and helped me in my PhD journey.

Words cannot express my gratitude to my family for their constant support. I would like to express my deepest gratitude to my parents, grandmother and siblings for their patience and making this journey a rewarding and exciting one.

This thesis is dedicated to my late grandfather Acharya Sudarshan Dwivedi who introduced me to this field and for giving me a good head start.

---

I would like to end by paying my heartfelt thanks and prayers to the Almighty for his unbound love and grace.

- Vandana Dwivedi

# Preface

This thesis investigates the machine learning based approach for analyzing the Hindi compound noun semantics. Noun compounds are interesting constructions from the perspective of theoretical linguistics and computational linguistics for the complex semantic relations between their constituents. The meaning of Compound Noun is composed of the meanings of the individual constituents and the way they are semantically related. Noun compound interpretation is the task of detecting this underlying semantic relation. For instance, a kitchen knife is a knife which is used in the kitchen (used in or purpose relation) whereas a steel knife is a knife made of steel (made of or constituent relation). This work explored different machine learning techniques as well as a knowledge base method which can predict the semantic relations between the constituents of compound nouns.

We developed a semantic relation set to annotate the semantic relations between the constituents of compound nouns, describing in detail the motivation for the relation set and the development process. Using the semantic relation set an annotated dataset is created to use as an experiment dataset with different machine learning algorithms. The inter-annotator agreements result indicate the reliability of the semantic relation and dataset for Hindi Compound Nouns analysis.

We treated compound noun interpretation problems as a classification problem for ML tasks and experimented with SVM and Random forest. Hindi WordNet is used as a linguistic resource and has got a significant output. We also experimented with Embeddings, since embedding captures more semantic features in language models and gives better results.

---

In this work, we also attempted to create a knowledge base in the domain of Ayurveda using the Generative Lexicon framework of lexical knowledge representation. Ayurveda text has a very limited corpus. Therefore, using a probabilistic model does not work. We developed a lexical knowledge representation database of some selected frequent nouns of Ayurveda corpus using the Generative Lexicon framework. The GL model is able to represent the natural polysemy of the word in the lexicon which is used for disambiguation tasks. Language models with rich knowledge encoded resources can be beneficial for developing linguistically precise probabilistic models.

# Contents

Certificate	iii
Declaration	v
Copyright Transfer Certificate	vii
Acknowledgments	ix
Preface	xi
Contents	xiii
List of Figures	xvii
List of Tables	xix
Abbreviations	xxi
Symbols	xxiii
<b>1 Introduction</b>	<b>1</b>
1.1 What is a Compound Noun? . . . . .	1
1.2 Why are they interesting? . . . . .	2
1.3 Classification of the Compound Nouns . . . . .	4
1.3.1 Paninian Linguistic tradition . . . . .	4
1.3.2 Western Linguistic tradition . . . . .	6
1.4 Objective of the work . . . . .	10
1.5 How of the work . . . . .	11
1.6 Outline of the Thesis . . . . .	12
<b>2 Processing of Compound Nouns in NLP</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 Compound Nouns as a Multi-Word Expression . . . . .	16

2.3	Defining Compound Nouns in NLP . . . . .	17
2.4	On the Compound Noun Interpretation . . . . .	18
2.4.1	Major theories in Cognitive Psychology . . . . .	18
2.4.2	Representation of Compound Nouns in Linguistics . . . . .	20
2.5	English Compound Noun Interpretation: A brief review . . . . .	21
2.5.1	The theoretical work that influenced NLP research . . . . .	22
2.5.2	First Corpus based study of Compound nouns in NLP . . . . .	24
2.5.3	The other major works in NLP . . . . .	26
2.6	Compound noun extraction and Interpretation works in Indian languages . . . . .	31
2.7	Summary and discussion . . . . .	35
2.8	Conclusion . . . . .	40
<b>3</b>	<b>Hindi Compound Noun Semantics: Creation of Dataset and Annotation Schema</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	Creation of Hindi Compound Noun Data . . . . .	45
3.2.1	Compound Noun data from TDIL . . . . .	45
3.2.2	Compound Noun data from Multi word expression list developed at IIT Bombay . . . . .	48
3.3	Curation of semantic relations set . . . . .	49
3.3.1	Samaas Based Classification of Hindi Compound Noun Data . . . . .	49
3.3.2	Procedure for Preparation of Semantic Relations . . . . .	51
3.3.3	Semantic relations for Hindi compound noun annotation . . . . .	54
3.4	Semantic relations in Domain Specific vs Domain Independent Compound nouns . . . . .	57
3.5	Evaluating the Compound Noun Dataset . . . . .	58
3.5.1	Inter annotator agreement . . . . .	59
3.5.2	Data and Semantic Relation set for the inter-annotator agreement experiment . . . . .	60
3.5.3	Evaluation Procedure . . . . .	60
3.5.4	Evaluation Method . . . . .	61
3.5.5	Result and analysis . . . . .	62
3.5.6	Prior work and Discussion . . . . .	63
3.6	Conclusion . . . . .	65
<b>4</b>	<b>Hindi Compound Noun Interpretation Using Machine Learning</b>	<b>67</b>
4.1	Introduction . . . . .	67
4.2	Machine Learning and NLP . . . . .	69
4.3	Theoretical Background of Machine Learning . . . . .	69
4.3.1	Performance metrics . . . . .	70
4.3.1.1	Confusion Matrix . . . . .	70

4.3.2	Classification with Support Vector Machines . . . . .	72
4.3.3	Classification with Random Forest Classifier . . . . .	73
4.3.4	Classification with BERT Classifier . . . . .	73
4.4	Word Embeddings . . . . .	74
4.4.1	Word2Vec Embedding . . . . .	76
4.4.2	BERT Embedding . . . . .	77
4.5	Compound Interpretation Dataset . . . . .	77
4.5.1	Computation of semantic relations . . . . .	78
4.5.1.1	Computational model for semantic relations identification . . . . .	78
4.5.2	Computation of Lexicalized, Name and Dvandva relation . . . . .	79
4.5.3	The compound noun dataset used with the classifiers . . . . .	81
4.6	Pre Processing of Data . . . . .	82
4.7	The Classification of semantic relations using Noun features and Machine Learning Classifier . . . . .	83
4.7.1	The Classification of semantic relations using Noun features . . . . .	83
4.7.2	Noun Features . . . . .	84
4.7.3	Experiments . . . . .	84
4.7.4	Results and Discussion . . . . .	86
4.8	The Classification of semantic relations using Word Embedding and SVM and BERT Classifier . . . . .	88
4.8.1	The Experiment 1: Word2Vec Embeddings and SVM classifier . . . . .	89
4.8.2	The Experiment 2: BERT Embeddings and BERT Classifier . . . . .	90
4.8.3	Result obtained From SVM classifier . . . . .	90
4.8.4	Result from BERT classifier . . . . .	91
4.9	Results and discussion . . . . .	92
4.10	Conclusion . . . . .	93
<b>5</b>	<b>Building a Knowledge Base Using Generative Lexicon for Domain-Specific Compound Nouns Interpretation</b> . . . . .	<b>95</b>
5.1	Introduction . . . . .	95
5.2	Introduction to Generative Lexicon theory . . . . .	96
5.3	Previous works with lexical knowledge bases . . . . .	100
5.4	Methodology of lexical knowledge base creation . . . . .	103
5.4.1	Creation of Corpus . . . . .	104
5.4.2	Annotation of Corpus . . . . .	106
5.5	Generative Lexicon and Semantics of Compound nouns . . . . .	111
5.5.1	Qualia Representation and description of <i>rasaayana</i> . . . . .	112
5.5.2	Qualia Structure of the word <i>rasaayana</i> and some compounds with it . . . . .	114
5.5.3	Rules for identification of the relation in <i>rasaayana</i> compounds . . . . .	117
5.6	Interpretation and Discussion . . . . .	120

5.7	Conclusion . . . . .	122
<b>6</b>	<b>Conclusion and Future Work</b>	<b>125</b>
6.1	Introduction . . . . .	125
6.2	Summarization of the thesis . . . . .	126
6.3	Limitations of the work . . . . .	128
6.4	Future direction of research . . . . .	129
	<b>BIBLIOGRAPHY</b>	<b>131</b>
	<b>A LIST OF PUBLICATIONS</b>	<b>147</b>
	<b>B HINDI HEALTH COMPOUND NOUN DATA</b>	<b>149</b>
	<b>C COMPOUND NOUN DATASET</b>	<b>157</b>
	<b>D QUALIA REPRESENTATION</b>	<b>169</b>

# List of Figures

1	ITRANS SCHEME . . . . .	xxv
2.1	Classification of Literature Review . . . . .	36
3.1	An example sentence from the corpus . . . . .	46
3.2	example sentences pattern for extracting the compound noun from the corpus . . . . .	47
3.3	Rosario Relation Set . . . . .	52
3.4	Tratz and Hovy Relations . . . . .	53
3.5	Semantic relations frequency distribution in our data . . . . .	57
3.6	Semantic Relation annotation comparison from the annotators . . . . .	63
3.7	Comparison of Agreement results between the annotators . . . . .	64
4.1	The computational model for compound noun interpretation . . . . .	79
4.2	Distribution of compound nouns into different semantic relations . . . . .	80
4.3	Distribution of frequent semantic relations with their frequency . . . . .	82
4.4	Binary Class classification result comparison using SVM and Random Forest(F1 score) . . . . .	87
4.5	Result from BERT classifier for each semantic relation . . . . .	92
5.1	GL Structure of a Word . . . . .	99
5.2	Frequency Distribution of Qualia relation in Compound noun data . . . . .	110
5.3	Relations and Qualia roles in the compound words with ‘rasaayana’ as head . . . . .	113
5.4	Relations and Qualia roles in the compound words with ‘rasaayana’ as head . . . . .	114
5.5	GL representation of the word <i>rasaayana</i> . . . . .	115
5.6	GL structure of the word <i>aahaara rasaayana</i> or <i>rasaayana</i> adminis- tered through food . . . . .	116
5.7	GL structure of the word <i>kuuTipravesika rasaayana</i> or <i>rasaayanaa</i> administered inside a house . . . . .	116
5.8	GL structure of the word <i>kharaliyarasaayana</i> ‘the <i>rasaayana</i> product made with <i>kharal</i> or mortar and pestle . . . . .	116

5.9	GL structure of the word <i>poTTalirasaayana</i> ‘the rasaayana product made in poTTali or cloth’ . . . . .	117
5.10	GL structure of the word <i>medhyarasaayana</i> ‘ rasaayana made for increasing intellect’ . . . . .	117
5.11	GL structure of the word <i>cikitsaarasaayana</i> ‘rasaayana for specific treatment’ . . . . .	118
5.12	GL structure of the word <i>aamlakiirasaayana</i> or rasaayana made of Indian gooseberry’ . . . . .	118
5.13	GL structure of the word <i>parpatirasaayana</i> ‘thin flake like medicine rasaayana’ . . . . .	118
5.14	Rule for representing Modifier Relation . . . . .	119
5.15	Rule for representing Purpose Relation . . . . .	119
5.16	Rule for representing Formal Relation . . . . .	120
5.17	Rule for representing Constitutive Relation . . . . .	120
B.1	Compound Noun Data Extracted from Hindi Disease Corpus . . . . .	150
B.2	Compound Noun Data Extracted from Hindi Disease Corpus . . . . .	151
B.3	Compound Noun Data Extracted from Hindi Disease Corpus . . . . .	152
B.4	Compound Noun Data Extracted from Hindi Disease Corpus . . . . .	153
B.5	Compound Noun Data Extracted from Hindi Disease Corpus . . . . .	154
B.6	Compound Noun Data Extracted from Hindi Disease Corpus . . . . .	155
C.1	WX Notation . . . . .	158
D.1	Qualia Representation of ‘rasa’ . . . . .	169
D.2	Qualia Representation of ‘dravya’ . . . . .	170
D.3	Qualia Representation of ‘prakriti’ . . . . .	170
D.4	Qualia Representation of ‘ghrita’ . . . . .	171
D.5	Qualia Representation of ‘curNa’ . . . . .	171
D.6	Qualia Representation of ‘vati’ . . . . .	172

# List of Tables

1.1	Classification of Sanskrit compounds . . . . .	6
1.2	Classification of compound nouns based on the semantic head . . . . .	7
1.3	Classification of compound nouns based on the parts of speech of constituents . . . . .	8
1.4	Classification of compound nouns according to Dressler (2006) . . . . .	9
2.1	Sinha’s multiword expressions list . . . . .	32
2.2	Distribution of Sanskrit Compounds Kulkarni et al (2010) . . . . .	33
2.3	Analysis of Hindi data for various types of paraphrases . . . . .	34
2.4	Summary table of major works . . . . .	38
3.1	Semantic Relation set proposed by us . . . . .	56
3.2	Inter annotator agreement result . . . . .	62
3.3	Comparison of Our Kappa Coefficient to other works . . . . .	64
4.1	Confusion Matrix . . . . .	71
4.2	Semantic Relation . . . . .	81
4.3	Result from SVM and Random Forest for multi class classifier . . . . .	85
4.4	Confusion matrix of all three relations classification results . . . . .	85
4.5	Result for SVM classifier . . . . .	91
4.6	F1 score for BERT classifier . . . . .	91
4.7	Comparison of SOTA results to our results . . . . .	93
5.1	Frequent Compound Nouns found in Ayurveda Corpus . . . . .	109



# Abbreviations

<b>AI</b>	Artificial Intelligence
<b>BERT</b>	Bidirectional Encoder Representations from Transformers
<b>RF</b>	Random Forest
<b>NLP</b>	Natural Language Processing
<b>SVM</b>	Support Vector Machine
<b>GL</b>	Generative Lexicon
<b>ML</b>	Machine Learning
<b>TP</b>	True Positive
<b>TN</b>	True Negative
<b>FP</b>	False Positive
<b>FN</b>	False Negative
<b>CN</b>	Compound Noun
<b>MWE</b>	Multi Word Expression



# Symbols

$w$  word

$N$  Noun

$k$  Kohen Kappa



ITRANS Transliteration Scheme

Vowels		Consonants	
ITRANS	DEVANAGARI	ITRANS	DEVANAGARI
a	अ	k	क
aa	आ	kh	ख
i	इ	g	ग
ii	ई	gh	घ
u	उ	~N	ङ
uu	ऊ	ch	च
RRi	ऋ	Ch	छ
RRI	ॠ	j	ज
e	ए	jh	झ
ai	ऐ	~n	ञ
o	ओ	T	ट
au	औ	Th	ठ
aM	अं	D	ड
aH	अः	Dh	ढ
		N	ण
		t	त
		th	थ
		d	द
		dh	ध
		n	न
		p	प
		ph	फ
		b	ब
		bh	भ
		m	म
		y	य
		r	र
		l	ल
		v	व
		sh	श
		Sh	ष
		s	स
		h	ह
		x	क्ष
		GY	ज्ञ
		shr	श्र

FIGURE 1: ITRANS SCHEME