# 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

**1 Introduction** 1
- 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

**2 Processing of Compound Nouns in NLP** 15
- 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

**3 Hindi Compound Noun Semantics: Creation of Dataset and Annotation Schema 43**
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

**4 Hindi Compound Noun Interpretation Using Machine Learning 67**
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

**5 Building a Knowledge Base Using Generative Lexicon for Domain-Specific Compound Nouns Interpretation 95**
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 *rasaayana* . . . . . 112
5.5.2 Qualia Structure of the word *rasaayana* and some compounds with it . . . . . 114
5.5.3 Rules for identification of the relation in rasaayana compounds 117
5.6 Interpretation and Discussion . . . . . 120

5.7 Conclusion . . . . . 122

**6 Conclusion and Future Work** **125**
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

**BIBLIOGRAPHY** **131**

**A LIST OF PUBLICATIONS** **147**

**B HINDI HEALTH COMPOUND NOUN DATA** **149**

**C COMPOUND NOUN DATASET** **157**

**D QUALIA REPRESENTATION** **169**