Chapter 1

Introduction

1.1 What is a Compound Noun?

The compound nouns are combinations of two or more nouns to form a new word by a process of word formation. The process of compounding is found extensively productive for deriving new words across the languages of the world. More explicit definitions of the compounds are theory-dependent as well as vary cross-linguistically. The prototypical compounds are composed of two prototypical lexical words and it does not include any clitic or a derivational morpheme Dressler. For instance, the English *class room* is an example of a prototypical compound, which is formed with two lexemes, *class* and *room*. Dressler also pointed out that derivation from a compound is not a compound but a derivational word as the English word *highlander* which has two lexemes *high* and *land* to make a compound plus a derivational suffix –er. Prototypically the members of the compounds are free morphemes with the exception of cranberry words where the first morpheme does not exist independently. Compound words generally consist of major lexical word classes, viz., noun, adjective and verb. However, the minor grammatical classes like indeclinable, preposition etc.

may form compound words. In this present work we are concerned with mainly noun-noun compounds.

1.2 Why are they interesting?

The compound words have generated interest among the theoretical linguists for their special semantic characteristics. The meaning of the compound word is not always predictable from the combinatorial meaning of the two words. One of the words in a compound is known as the head word and it carries the main semantic character of the word. How this head word is related to the other words plays an important role for understanding the underlying sense of the resulting compound. The classification of the compounds on the basis of the semantics of the constituents has been an important topic in theoretical linguistics right from the Sanskrit grammatical tradition to the modern structural and generative linguistics. The polysemy of the compounds has also attracted attention of the scholars. Dressler (2006) quoting Coseriu (1975) cites an example of a German compound, strabe-n-verka¨uf-er lit. 'street seller' which may mean a builder who sells streets to a third world country but generally it actually means a seller who sells something on the street. There are other problems with the interpretation of compounds with the same head word when the other word changes. A *gas pipe* is a pipe to supply gas but a *steel pipe* is made of steel.

This feature of compound noun semantics has actually created a lot of interest among the psycholinguists from their processing point of view. A number of psycholinguistic theories attempted to explain how human mind represents and processes compound words. A large number of works in psycholinguistics concentrated on understanding how concepts are combined to make a new word, especially a new compound and what role context plays in the interpretation of that compound F. J. Costello;

F. Costello & Keane; F. J. Costello & Keane; Gagné; Gagné; Wisniewski; Wisniewski & Love

When the Natural Language Processing (NLP) research started in English, one of the areas of semantics which posed a huge challenge to the computer scientists is the Multi word Expressions. These are expressions containing more than one word and when combined, they produce a non-combinatorial meaning. The compound words fall in this category. The automatic semantic interpretation of different English compounds has long been an area of research in computational linguistics. Noun compounds are interesting constructions from the perspective of the computational linguistics for the complex relations between their constituents. Interpretation of noun compounds is the task of uncovering a relationship between component nouns of a noun compound. The meaning of a noun compound 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 for or purpose relation) whereas a *steel knife* is a knife made of steel (made of or constituent relation).

The problem entered the research arena of NLP of Indian languages in recent times when we have moderately managed the basic morpho-syntactic problems of a language. As we are trying to develop systems with more knowledge of semantics for various applications in the field of applied Artificial Intelligence, we are facing more the challenge of handling MWEs. Identifying and interpreting compound words (including complex predicates and compound nouns) is a challenging task in Indian languages.

The present work is set up in this background of research on compound nouns. Indian languages use compounding as a very productive word formation strategy. It is considered one of the most important areal features of 'India as a linguistic area'. Computation of compound noun semantics for Indian languages becomes a

major challenge for further development of many other text processing works which heavily rely on semantics such as text summarization, building automatic dialog system, machine translation, information extraction etc. This study attempts to fill this gap in the computation of compound nouns semantics for Hindi.

1.3 Classification of the Compound Nouns

In this section we discuss two major theoretical linguistic paradigms in which compound nouns have got a major attention. One is the Paninian Linguistic tradition from ancient India and the other is the Generative Grammar tradition in the modern west. Both these theoretical paradigms considered compounding as a major word formation strategy in any language.

1.3.1 Paninian Linguistic tradition

Panini was an ancient Indian grammarian of circa 5th century BC who wrote an excellent work on Sanskrit grammar consisting of eight chapters and named *Ashtadhyaayii*. This grammar had a specific chapter on *samaasa* or compound words which talks about its classification and semantics. The compounds of Sanskrit are classified into four major types depending on the relative prominence of their constituents in the meaning formation of the composed word. The basic types are: *avyayiibhaava, dvandva, tatpurusha* and *bahuvriihi*. The *tatpurusha* has further subcategories: *karmadhaaraya* and *dvigu*.

The *avyayiibhava* is the compound where the first constituent is semantically more prominent in the compound word and the first component of this compound is always an *avyaaya* or indeclinable word. E.g. *pratyeka* 'every' consists of two words *prati* 'each' and *eka* 'one'. The *dvandva* has two or more constituents and all are equally

important to provide the meaning of the compound nouns. E.g. *raatridina* 'day and night' consists of two words *raatri* 'night' and *dina* 'day'.

Tatpurusha is the compound where the latter word is semantically important. The former word is related to the latter by different case marking and the sub-classes under *tatpurusha* compound are named according to this case marking. These sub-classes are *dvitiiyaa* (2nd), *trtiiyaa* (3rd), *caturthii* (4th), *panchamii* (5th), *shashthii* (6th) and *saptamii* (7th) tatpurusha. E.g. *raajaputra* 'king's son' consists of two nouns *raajna* 'of the king' and *putra* 'son' and the case marker on the first word is a 6th case ending. Therefore, specifically it is a *shashthii tatpurusha* compound.

Bahuvriihi is the compound where neither of the constituents is semantically important and the resulting compound gets a distinct meaning. E.g. *pancaanana* 'one who has five faces, *Shiva* in particular' consists of two words *pancha* 'five' and *aanana* 'face'.

The *karmadharaya* is the compound where the first word is a modifier of the second word and the meaning of the second word is prominent in the compound. E.g. *niilkamala* 'blue lotus' consists of two words *niila* 'blue' and *kamala* 'lotus'.

Dvigu is a special case of *karmadhaaraya* where the first word is a number and the second word is semantically prominent in the compound semantics. E.g. *trinayana* 'three eyes' consists of two words *tri* 'three' and *nayana* 'eye'.

The rule of Sanskrit compounding is when two words are combined, the case marker of the first word is deleted. However, there are some exceptions to this rule and they are known as *aluk* (literally not deleted) *samaasa*. The table 1.1 shows the different types of compounds with examples.

Compound Noun Type	Example
Avyayiibhaava	*pratyeka* (prati 'each' and eka 'one') "every"
Tatpurusha	raajaputra (raajna 'of the king' and putra 'son') "king's son'"
Karmadharaya	*niilkamala* (niila 'blue' and kamala 'lotus')'blue lotus'
Dvigu	*trinayana* (tri 'three' and nayana 'eye')'three eyes'
Bahuvriihi	*pancaanana* (pancha 'five' and aanana 'face'.) "one who has five faces, *Shiva* in particular"
Dvandva	*maata-piita* (mother-father) "Parents"

TABLE 1.1: Classification of Sanskrit compounds

1.3.2 Western Linguistic tradition

Generative linguists of the 20th century have discussed compounding as a productive word formation strategy. One of the criteria used in theoretical linguistics for classification of compound nouns is headedness feature. Morphological words are also assumed to have heads like the head of a syntactic construction. According to the headedness (semantic head) of the compound nouns are divided into endocentric and exocentric. Endocentric compounds are the compounds which have heads inside the compound as in *apple juice* which is a type of juice and the head of the compound is juice. The core meaning of the compound comes from the meaning of the head and compound characteristics also follow the head characteristics. The Compound without a head or when the head is outside the compound is called an exocentric compound. English has very few such compounds. The Sanskrit equivalent of exocentric compound is *bahuvriihi*. The difference between these two types of compounds sometimes depends on the interpretation factor, like a *blackbird* is a species of bird which can be green in color or white in color or black in color. The

compound *blackbird* is endocentric or exocentric depending on which bird is actually referred to. Head of the compound is generally the rightmost element in the compound except for some language like French. Exocentric compounds also derive their syntactic and semantic characteristics of the rightmost (head) element. Only the semantic head deviates in the exocentric compound nouns.

Coordinate compound is the third type of compound where both the constituents share head-like characteristics and the meaning comes equally from both the constituents. Other name of the compound is copulative or *dvandva* in Sanskrit. Like *mother-father* which means parents or *actor friend* whose meaning is a person who is both a friend and an actor. The table 1.2 shows them with example.

Compound Types	**Example**
Endocentric	school bus (bus is the head)
Exocentric	blackbird (neither of the words is head)
Coordinate	actor friend (both are heads)

TABLE 1.2: Classification of compound nouns based on the semantic head

Noun-noun compounds are "a salient feature of Sanskrit and Germanic word formation, whereas in other Indo-European language families like Latin, Indian and the Romance languages, Celtic or Slavic, they are marginal and replaced by syntactic phrases" KASTOVSKY. The equivalent pattern for the compound nouns in these languages follow noun-preposition-noun construction. Modern Indian languages sometimes also follow the same construction. In Indian languages, there is a category of compound nouns which has a genitive marker in between the constituents and most of the compound nouns can be written as the form of N-Gen-N construction. As *lohe kii kursii (iron chair)* and *lakdii kaa makaan* (wooden house). English compound nouns can be of different types based on the structure. The table 1.3 shows them with examples.

Compound structure	Example
N + N	iron man, pain killer
V + N	swimming pool, washing machine
P + N	after life, back date
N-pre-N	mother-in-law

TABLE 1.3: Classification of compound nouns based on the parts of speech of constituents

There are other types of classification of compounds based on the transparency criterion. The meaning of the compound is understood by the meaning of the constituents of the compound along with their composite meaning. However, this is not in the strict Fregean sense of compositionality of meaning construction. With the varying degree of composition there are different types of compound nouns compositional to lexicalized. Lexicalized compounds are the compounds which cannot be formed by the productive rules and must be placed in a dictionary like *brahamani chiil* is a type of eagle in Hindi.

Deictic compounds, like other categories of deixis, cannot be interpreted without contextual information as they are normally created in conversational settings "to satisfy a fleeting discourse need"RYDER (1994). Deictic compounds are often excluded from linguistic studies of compound interpretation because their meaning depends on extra-linguistic information; for example, DOWNING (1977)uses the term deictic compounds to describe the subset of "novel compounds used in conversational situations" which she excludes in her study. In this context, DOWNING (1977) introduces the famous example of deictic compounds, *apple-juice seat*, which is often cited in NLP studies to demonstrate the difficulty of interpreting noun–noun compounds. Novel compounds, in contrast, are interpretable—to some extent—out of their original context. However, as Ryder explains, "this requirement still does not necessarily limit the kinds of relationships that could hold between the element nouns" RYDER (1994). Furthermore, novel compounds refer to potentially more

permanent states, in contrast to deictic compounds that are generally derived from temporary states (e.g. the apple-juice placed in front of the seat). To illustrate the limited interpretability of novel compounds, Ryder (1994) explains that the hearer only has access to the predictable meaning of pencil sharpener (a tool for sharpening pencils), but not its size or shape. Established compounds are originally novel compounds, but they have been accepted by the language speakers and become part of their lexicon over time. This does not imply that all established compounds are idiomatic, but idiomatic compounds have to be established, otherwise it would be impossible to interpret them. In fact, Ryder (1994) points out that established compounds become like words and undergo semantic drift. She gives the example of cod fish, which is the common name for a fish species, but according to Ryder *cod fish* originally meant "a fish that is like a cod". Based on Ryder's discussion of the three types of noun–noun combinations, it is plausible to assume that the boundaries between these three groups are not always clear-cut.

Compound Types	**Example**
Transparency of both the members	school bus
Transparency of the head member	straw-berry
Transparency of the non-head member	jail-bird
Opacity of both the members	hum-bug

Table 1.4: Classification of compound nouns according to Dressler (2006)

Following Libben (1998), Dressler (2006) categorize compounds on the basis of the morpho-semantic transparency of the constituents of the compound. The transparency of the head member is more important than the non-head members for the interpretation of the compound nouns. The table 1.4 arranges them according to the transparency and the level of difficulty in processing, the most transparent and the least difficult placed at the top. The transparency of the head member is more important than the non-head member for processing of the compounds.

The above discussion on the classification of the compound words in theoretical linguistics reveal a fact there is not one single criterion of classification. Some classification is based on the headedness, whereas other is based on the semantic transparency of the constituent words. Some other classification even includes the context of creation of the compound words. But no criterion captures the semantic relation between the head and the non-head member which is important for the interpretation of the compound word.

1.4 Objective of the work

The primary objective of the work is to classify, interpret and analyze the compound nouns of Hindi. The theoretical linguistics, right from the Indian Paninian grammatical tradition to the western generative linguistics, has placed a lot of importance on the classification of the compound nouns. But it failed to create a suitable scheme of classification that can be adopted for the automatic interpretation of compound nouns. The reason being, the Paninian grammatical tradition as well as the western linguistics mainly focused either upon the relative semantic prominence of the constituents of a compound when they form a new word or the classification is based on the headedness property. Neither of these theories focuses on the internal relation of the constituent words. For instance, both *kitchen knife* and *steel knife* are *tatpurusha* compounds in the Paninian tradition which is defined as the compound where the semantics of the second word is prominent in the resulting compound word. Both the compounds are endocentric compounds according to the western scholars Smith (1991); Williams (1981) as the head word is inside the compound. Most of the English compounds are actually endocentric compounds and the head word is on the right. Therefore, such a classification scheme is not able to distinguish between these two compound nouns. In our data also we found mostly endocentric compounds which are of tatpurusha type in the Paninian tradition. The major research issues and questions addressed in this dissertation are the following.

- To identify and annotate different semantic relations in the compound nouns of Hindi.
- To make a suitable optimal schema of classification for Hindi that is able to capture the Hindi specific compounds.
- To find out to what extent machine learning algorithms can be applied to Hindi data to get a considerably good result of classification.
- Which classification algorithms are most suitable for doing this classification job?
- What kind of approach suits best for a low resource language like Hindi for doing such a task?

1.5 How of the work

We handled these issues briefly in the following ways.

- We used three different datasets for this work. The first dataset is extracted and made by us from the annotated corpus of the health domain available in the website of Technology Development for Indian Languages (TDIL). The number of noun compounds found here are only 200. We modified some existing English compound noun classification schema and gave it an optimal size by deleting some English tags and adding some new tags suitable for Hindi. The number of relations in the relation set is a total of 15. Initially, we annotated this data of 200 noun compounds using these 15 relations.
- The second data set used by us is a dataset of Multiword Expressions which was publicly available and made by IIT Bombay from a general domain corpus. We extracted only the Noun+Noun compounds from this consisting of a total

1500 words. We annotated this data set with the 20 semantic relation set (the set consists of the same set as of our first work and also some more relations which we observed in the general domain compound noun data) to check the suitability of those 20 relations.

- To find out the suitability of the machine learning algorithms, we experimented with Random Forest and SVM in the beginning and then used Word2Vec embedding as a feature in a BERT model. We found that our data size is small and some relations are more frequent than others, therefore, the data set is imbalanced. So, we did these experiments to identify three major frequent relations and a mixed bag consisting of all other infrequent relations.

- While doing semantic interpretation of compound nouns we realized that Hindi does not have good lexical and knowledge resources. We used Hindi WordNet for excluding some lexicalized relations from our data set. However, for capturing the complex relations between two nouns in a compound noun word, only WordNet is not enough. We felt that machine learning algorithms along with a strong support of a lexical Knowledge base is the most suitable approach for getting a good result in any semantic interpretation task.

- In order to build a domain-specific knowledge base, we took the data from the Ayurveda domain. Our first data was also from the health domain, so we decided to extend that in the Ayurveda, the ancient Indian health science. We used Generative Lexicon as a framework for knowledge representation in this work for the Ayurveda compound nouns.

1.6 Outline of the Thesis

The rest of the thesis is organized into five chapters. The outline of the thesis is as follows:

CHAPTER 2 discusses the state of the art research on the noun compounds in general as done in the theoretical linguistics, psycholinguistics and NLP domains. It also provides some key concepts, problems and terminology related to compound noun processing.

CHAPTER 3 introduces the data set created and proposes a taxonomy of semantic relation for compound noun interpretation for Hindi based on existing different state of the art taxonomies. We then calculated inter annotator agreement to check the homogeneity of relations and developed an annotation schema. We also reviewed compound noun bracketing for the compound nouns having more than two noun constituents. We also review samaasa based compound noun interpretation as well as other linguistic framework free paraphrasing and prepositional paraphrasing for representing the meaning of compound nouns. This chapter also describes the dataset curation for our further experiments using machine learning for Hindi compound noun interpretation.

CHAPTER 4 focuses on the automatic compound noun interpretation using different machine learning algorithms and discusses and compares the results from these algorithms.

CHAPTER 5 introduces Generative Lexicon as a Lexical Knowledge Representation Model and uses it for representing some most frequent nouns used in the compound words in Ayurveda domain.

CHAPTER 6 the final chapter summarizes the work and discusses its limitations. It also identifies some future research issues related to the computation of compound noun semantics.