Semantic Processing of Compounds in Indian Languages

Amba Kulkarni[1] Soma Paul[2] Malhar Kulkarni[3]
Anil Kumar[1] Nitesh Surtani[2]

1. Department of Sanskrit Studies, University of Hyderabad
2. International Institute of Information Technology, Hyderabad
3. School of Humanities and Social Sciences, I.I.T. Bombay

apksh@uohyd.ernet.in, soma@iiit.ac.in, malharku@gmail.com,
anil.lalit22@gmail.com, nitesh.surtani0606@gmail.com

Abstract

Compounds occur very frequently in Indian Languages. There are no strict orthographic conventions for compounds in modern Indian Languages. In this paper, Sanskrit compounding system is examined thoroughly and the insight gained from the Sanskrit grammar is applied for the analysis of compounds in Hindi and Marathi. It is interesting to note that compounding in Hindi deviates from that in Sanskrit in two aspects. The data analysed for Hindi does not contain any instance of Bahuvrīhi (exo-centric) compound. Second, Hindi data presents many cases where quite a lot of compounds require a verb as well as vibhakti(a case marker) for its paraphrasing. Compounds requiring a verb for paraphrasing are termed as madhyama-pada-lopī in Sanskrit, and they are found to be rare in Sanskrit.

Keywords: Multi Word Expression, Compounds, Pāṇini, Aṣṭādhyāyī, semantics.

1 Introduction

Noun compounds represent a linguistic device of encrypting information that makes their analysis a challenging NLP task. For example, consider a compound 'ballpoint pen'. It is made up of two nouns: a modifier 'ballpoint' and a head 'pen'. The modifier is again a compound that is composed of two nouns 'ball' and 'point'. The relation between 'ball' and 'point' and in turn between 'ballpoint' and 'pen' is not encoded anywhere on the surface although a user of the language can decodify the meaning correctly. The study of compounds involves two major tasks: 1) automatic identification and extraction of compounds from natural language texts and 2) syntactic and semantic analysis of compounds. The task of identification of compounds becomes significant because of orthographic vagaries. Orthographic conventions for writing compounds may vary from language to language and even within the same language. For example, in Sanskrit a compound is a single word, while in English (as exemplified above) as well as in modern Indian Languages (ILs) we find the following conventions of writing: components written with or without a space and components separated by a hyphen. When compounds are written without a space, the adjoining phonemes undergo euphonic changes as in *gaṅgā-udaka* changing to *gaṅgodaka*. Analysis of compounds primarily involves the expansion of these syntactically condensed constructs with an aim to unfold the meaning of the constructions (Butnariu and Veale, 2008; Girju et al., 2007; Kim and Baldwin, 2006; Kumar, 2012; Nakov, 2008; Nastase and Szpakowicz, 2009; Séaghdha and Copestake, 2007). Semantic analysis of compound is significant for various NLP applications including machine translation, information extraction and so on. Here we discuss one example to show how semantic analysis helps in machine translation. For example, let us consider the English compounds *cancer death* and *room temperature.* Lexical substitution of components of these compounds into Hindi would produce the following result: *kainsara mauta* and *kamarā tāpamāna* which are not legitimate constructions in Hindi. However, the semantic paraphrase of the two compounds, namely, *death from cancer* and *temperature of room* will be helpful for achieving the correct translations of the compounds: *kainsara se* (or *ke kāraṇa*) *mauta* and *kamare kā tāpamāna*. Currently there exist two different approaches in Computational Linguistics to deal with this phenomenon (Paul et al., 2010). They are (a) Labeling the semantics of compound with a set of abstract relations (Girju et al., 2003) and (b) Paraphrasing the compound in terms of syntactic constructs. Paraphrasing is again done in three ways: (i) with prepositions (*war story* vs *story about war*) (Lauer, 1995) (ii) with verb+preposition nexus (*war story* vs *story pertaining to war*, *noise pollution* vs *pollution caused by noise* (Finin, 1980) (iii) with Copula (*tuna fish* vs *fish that is tuna*) (Vanderwende, 1995).

Detailed study of Sanskrit compound processing had been taken up recently (Kumar, 2012), and the insights gained there were found useful for processing the compounds in ILs. After looking at the features of Sanskrit as described in the Sanskrit grammar, in the third section we describe the automatic Sanskrit compound processor, followed by the insights we gained from this processor to identify the compound tags and also the semantic categories necessary to carry out the compound analysis automatically. The fourth section discusses as a case study, use of these tags for Hindi and Marathi compound types. Conclusion follows in the fifth section.

2 Sanskrit Compounds

Pāṇini (500 BCE approximately) has described the process of compound formation in Sanskrit in his grammar called Aṣṭādhyāyī. He used the term Samāsa for Compound. The word Samāsa literally means "Throwing out together" which in the context means "throwing out the words together”. This concept implies that words are thrown out of mouth by human beings in syntactically related structures. There are three main features of a Compound: 1. One word (*aikapadya*), 2. One meaning (*aikārthya*) and 3. One accent (*aikasvarya*). According to Pāṇini, two words within a sentence can form a compound if and only if there is syntactic compatibility (sāmarthya) amongst them. Mere adjacency of words does not allow them to get compounded. The words should be first syntactically related and further should possess the quality of being used by the native speakers as one unit. Further, what distinguishes a compound from a non compounded word group within a sentence are various morphological as well as syntactical features such as a) loss of case, b) absence of intervention of other words, c) no possibility of relation of a non-head word within a compound with another word in the sentence, d) absence of expression of number of the first component.

There are broadly speaking four types of compounds in Sanskrit: 1. Avyayībhāva, 2. Tatpuruṣa, 3. Dvandva and 4. Bahuvrīhi. Semantically, Avyayībhāva and Tatpuruṣa are endocentric compounds with the head typically to the left and right respectively. Dvandva is a copulative compound while Bahuvrīhi is an exocentric compound. Many of the compounds are compositional and hence can be generated with the help of a rule base. However, there are some compounds which are non-compositional and they are treated separately in Pāṇini’s grammar. These compounds are called nitya samāsa (obligatory compounds), that is they are always used in compounding form. Such compounds either can not be paraphrased at all (*Avigraha*) or involve extra words other than the components for their paraphrase (*Asvapadavigraha*). Following two examples will illustrate this point:

i) *aśvakarṇa* (name of a medicinal plant) is a compound made up of two words, *aśva* ‘a horse’ and *karṇa* ‘an ear’. When both these words are joined together, they result into a compound, *aśvakarṇa*; but, this compound has no traces of the meanings of it’s components. Therefore, this compound can not be paraphrased in terms of its components.

ii) *Kākapeyā* (meaning: a river which contains water potable only by crows). This compound is made up of two components, namely, *kāka* ‘a crow’ and *peya* ‘potable’. As we see, there is an additional semantic element of censure which goes to make this compound. This additional meaning is obviously not a part of the component meanings.

The meaning of compounds may or may not be compositional. Based on the discussions in traditional Sanskrit grammar sources, we see a spectrum of compositionality as illustrated below (Shastri, 2006).
i) purely compositional (*sambaddhārtha*)
ex: *rāja-puruṣaḥ*
gloss: King - man
meaning: King’s man

ii) compositionality with fixity of expression (*samprekṣita*)
ex: *khaṭvā-āruḍhaḥ*
gloss: Bed - one who climbs
meaning: One who climbs the bed without completing the education

iii) Non compositionality (with some predictability) (*samgatārtha*)
ex: *citra-gu*
gloss: colorful - cow
meaning: One who has a colorful cow

iv) Non compositionality (*samsṛṣṭārtha*)
ex: *aṣva-karṇa*
gloss: horse - ear
meaning: Name of a medicinal plant.

In Sanskrit compounds are always written without any space in between. But modern Indian languages do allow space in between. Therefore, this spectrum of meanings makes it difficult to decide whether a group of words written with space in between is a compound or not, making the identification of a compound a challenging task for these languages.

3 Sanskrit Compound Processing

Sanskrit is rich in compound formation. Almost every fifth or sixth word in a randomly chosen Sanskrit text is a compound. Compound formation being very productive, we can not list all the compounds in a dictionary. An automatic compound processor was developed (Kumar, 2012) as a part of Computational Toolkit for Sanskrit. This compound processor provides a general architecture for processing compounds.

Analysing a Sanskrit compound involves
i) Segmentation,
ii) Deciding the constituency structure,
iii) Identification of relations between the constituents, and
iv) paraphrasing it.

3.1 Segmentation

In Sanskrit a compound is always written without any space. Moreover, the phonemes of the adjoining components necessarily undergo euphonic changes. Splitting involves reversing these euphonic changes. For instance, the compound *gaṅgodaka* is segmented as *gaṅgā-udaka*. It is possible that a word is ambiguous leading to multiple possible splitting. In Sanskrit the authors have taken advantage of this ambiguity which resulted in many texts with pun. The splitter should be able to produce all possible splits and also rank them if possible. A splitter needs sandhi rules and also a morphological analyser to validate the splits. Two different methods have been followed for building a Sanskrit splitter (Mittal, 2010). In the first approach FST built for morphological analyser is augmented with the sandhi rules (Huet, 2009). In the second approach a given string is split in all possible ways following the sandhi rules, and then the splits are validated through the morphological analyser. This is closer to the GENerate-CONstrain-EVALuate model of Optimality theory (Prince and Smolensky, 1993). The sandhi rules split the given string into all possible ways, then constraints, such as every component of the split should be a valid morph, are applied, and finally the possible splits are ranked using the language and split model. The results of this splitter are quite good, with 93% of cases the first split is correct. This method, though sounds good, practically ends up generating thousands of splits 90% of which are not validated morphologically. Thus this method is computationally less efficient. On the other hand a splitter built by augmenting the FST with sandhi rules is computationally very efficient, since it splits the string only if it is morphologically valid thereby avoiding unnecessary splits. If this FST is further augmented with a proper model for sandhi rules and the lexicon, better results are expected.

3.2 Constituency Parser

Constituency parser takes an output of the segmenter and produces a binary tree showing the syntactic composition of a compound corresponding to each of the possible segmentations. Each of these compositions shows the possible ways various segments can be grouped. To illustrate various possible parses that result from a single segmentation, consider the segmentation a-b-c of a compound. A compound being binary[1], the three components a-b-c may be grouped in two ways as *<a-< b-c>>* or *<<a-b>-c>*. Only one of the ways of grouping may be correct in a given context (unless the text has intended pun) as illustrated by the following two examples. Parse of *three-meter-wide* is *<<three-meter>-wide>,* and that of *iron water pump* is *<iron-<water-pump>>*. The number of possible parses increase exponentially as the number of components increase. The problem of constituency parsing is similar to the problem of completely parenthesizing *n+1* factors in all possible ways. Thus the total possible ways of parsing a compound with $n+1$ constituents is equal to a *Catalan number,* C_n (Huet, 2009). In the absence of any morpheme marking the relation between the components, the constituency structure is governed by the compatibility of meanings of the components involved. Hence to decide the constituent structure, a semantically rich lexicon is needed. In the absence of any such lexicon, the statistical properties of the manually tagged corpus were used (Kulkarni and Kumar, 2011) to decide the constituency structure. For compounds with 3 components, it had a F-measure of 93.66, in case of compounds with 4 components, the F-measure dropped to 65.4. The corpus had compounds with as many as 10 components. On average in 86.5% of cases, the machine could produce the correct parse. Pāṇini has also provided certain rules with morphological constraints on the final component[2], or certain group of words as an initial component[3]. These rules help in prioritising the grouping. These rules, though are written for Sanskrit, hold good across languages. Here is an example of a rule which deals with three components. The sūtra *(diksaṅkhye) taddhitārthottarapadasamāhāre ca (2.1.50)* says that in case of a compound with three components with number or direction indicating word as the first component, the first two components combine first. This holds good for other languages as well. For example, *one day cricket match* is *<<one-day>-<cricket-match >>* and *South Indian Association* is *<<South-Indian>-Association>*. However, since the surface forms for adjectival usage and compounding forms in English being the same, one may have ambiguous expressions such as *South sea route.* But in Marathi which distinguishes between the adjectival and compounding usage, *dakṣiṇa sumudrī mārga* (<<south-sea>-route>) is different from *dakṣiṇī sumudrī mārga* (<south-<sea-route>>).

3.3 Type Identifier

The semantic classification of compounds given by Pāṇini is not only restricted to Sanskrit language per se, but is more general. For example, the Cambridge Grammar of the English language (Huddleston and Pullum, 2002) uses this classification to describe compounds in English. *Water pump* is an endocentric compound. An endo-centric compound typically shows a hyponymic relation with the head noun. An *egghead* is a bahuvrīhi compound meaning a person whose head resembles the shape of an egg, i.e. a high forehead, and hence intellectual. An example of a coordinative compound is *Hewlett-Packard* meaning a company whose owners are both Hewlett and Packard, or *secretary-treasurer* which stands for a person who is both a

[1] With a possible exceptions of coordinative and certain other rare compounds.

[2] for example, non-finite verbs ending in 'kta' suffix.

[3] for example sūtras corresponding to the 'avyayībhāva' type.

secretary as well as a treasurer. The type of a compound thus is useful in deciding the meaning of a compound. In order to decide the type of a compound, an access to the semantic content of its constituents, and possibly even to the wider context is needed. Now the immediate question is whether this classification of compounds into four classes is sufficient, or do we need further sub-classification. The grammatical texts sub-divide the Tatpuruṣa compounds further into sub-classes based on the case marker the first component takes when the compound is paraphrased. As an illustration, a compound *vidyānipuṇa* 'one who is sharp in the studies', when paraphrased, the first component takes a locative case marker, another compound *Daśarathaputraḥ* 'son of Dasharatha' takes a genitive marker, while the compound *vyāghrabhaya* 'fear of a tiger' takes an ablative case marker in Sanskrit. Based on the paraphrase, a set of 56 fine-grain tags was identified (Kumar et al., 2009) for Sanskrit.

It is important to note the level of semantics the compound tags deal with. Consider the compounds *rājapuruṣaḥ* 'King's servant', *Daśarathaputraḥ* 'son of Dasharatha', and *vṛkṣaśākhā* 'branch of a tree'. In the first case the relation between *rājan* 'king' and *puruṣa* 'man' is that of servant-master (*sevya-sevaka*), in the second the relation between *Daśaratha* and *putraḥ* 'son' is of father-son (*pitā-putra*) and in the third case the relation between *vṛkṣa* 'tree' and *śākhā* 'branch' is part-of (*avayava-avayavi*). However, in all the three cases instead of specifying these deeper relations, relation between the components is expressed through the genitive case suffix in the paraphrase of these compounds as *rājñaḥ puruṣaḥ* 'King's servant', *Daśarathasya putraḥ* 'son of Dasharatha', and *vṛkṣasya śākhā* 'branch of a tree', and thus these are classified as *Ṣaṣṭhī-Tatpuruṣa* 'genitive endocentric compounds with head to the right'. In other words, the classification is not guided by the deeper semantics, but by the paraphrase of a given compound, or by what the language expresses. Thus, on the one hand, to decide the meaning of a compound, we need a fine-grain tagset, at the same time, it should not be as fine-grained as to distinguish between the meaning of genitive cases in *rājñaḥ puruṣaḥ, Daṣarathasya putraḥ* and *vṛkṣasya śākhā*.

Assuming that we follow the fine-grained classification of compounds as dictated by the paraphrase, the question is, to what extent is it possible to decide the relation between the words only on the basis of components involved? For classification, a manually tagged corpus was used as a training data. A corpus of size 800K is tagged manually for the compounds in context by the Sanskrit Consortium. This had 92K instances of compound words. The distribution of frequent compounds is given in Table 1.

Type	Percentage
Endocentric	58.70
Karmadhāraya (IS_A)	18.11
Exocentric	11.04
copulative	5.67

Table 1: Distribution of Sanskrit Compounds

The endocentric compounds were further classified on the basis of missing case marker in the paraphrase. It was found that 55% of these compounds required genitive case marker. (Kumar et al., 2010) reported that the precision of a statistical classifier on this data considering only the major classification, is 72.7%. Allowing sub-divisions resulting into fine-grained tagset lowers the performance to 63%. Statistical taggers perform well provided the training data is sufficient. So their performance goes down on compounds of rare type. Pāṇini has provided several sūtras in Aṣṭādhyāyī which deal with rare compound types. These sūtras provide various kind of semantic conditions under which a particular type of compound formation takes place. After going through the relevant sūtras, we observe that the conditions stated by Pāṇini fall under the following categories.

1. A restricted list of allowed components in certain type of compounds is provided.
2. A restriction in terms of special inflectional suffix / derivational suffix / category is mentioned.
3. A restriction is stated in terms of special technical terms, which are theory internal.
4. A restriction in terms of semantic relations between the components is mentioned.
5. Semantic property of the component is stated as a condition.

Out of these, the fourth and fifth category are important. The fourth category provides us clues for the important types of relations. Efforts such as Sanskrit WordNet (Kulkarni et al., 2010) or on marking semantic information in various kośas such as Amarakośa (Nair and Kulkarni, 2010) are concerned about lexical as well as semantic relations. In the sūtras related to compounds, we found the mention of following semantic relations.
i) *viśeṣaṇa-viśeṣya-bhāva* 'modifier-modified relation'
ii) *upamāna-upameya-bhāva* 'analogy or comparison'
iii) *avayava-avayavī-bhāva* 'part-whole relation'
iv) instrument-action relation.

The fifth category of conditions puts certain restrictions on the component in terms of semantic properties such as the component should be either a number or a direction or a color or a class indicating word or an adjective. This provides us a clue that the lexicon should have these semantic properties marked to enable automatic compound processing.

3.4 Paraphrase

The tagset for Sanskrit is dictated by the paraphrase. So except for some rare compound types with irregular paraphrases, typically each tag correspond to a well defined paraphrase (Kumar et al., 2009), distinct from the other, and one can then generate the paraphrase automatically. For example, the paraphrase rule for a *Ṣaṣṭhī Tatpuruśa* (T6) 'genitive compound' is given as below.

<x-y>T6 = x{6} y

where x{6} stands for the nominal form of x in genitive case. The paraphrase rules for the complete tagset along with the paraphrase generation is discussed in (Kumar et al., 2009). The problematic cases were those with an elision of certain terms called *madhyampadalopī*. For example, the paraphrase of *Śākapārthivaḥ* 'vegetable - human' is *Śākapriyaḥ pārthivaḥ* 'a human who likes vegetables'. In order to get this paraphrase, we need to insert appropriate content word. Such compounds are rare in Sanskrit, and are listed as exceptions.

3.5 Insights Gained

This detailed study of Sanskrit compounds has helped us in getting a good insight into the compound processing. To understand the meaning of a compound, or to translate a compound into another language, first one needs to understand the underlying constituency structure. For example,
South Indian Cricket Association = *<<South-Indian>-<Cricket-Association>>*,
South Indian Food Plaza = *<<<South-Indian>-Food>-Plaza>*, and
Colon Cancer Tumor Suppressor Protein = *<<<<Colon-Cancer>-Tumor>-Suppressor>-Protein>*.

The second important step in understanding of a compound is to identify the relation between the component pairs. These relations are classified broadly into four categories depending upon the position of the head in the compound. The four major types of compounds were further sub-divided taking into account the differences in their paraphrases. Manual tagging of the Sanskrit compounds revealed that only few of the compounds are very frequent. They include *Tatpuruṣa, Bahuvrīhi* and the *Karmadhāraya*. In case of *Tatpuruṣa* compounds, the paraphrase requires appropriate case marker which shows the relation between the components. Among the *Tatpuruṣa* compounds *Ṣaṣṭhī Tatpuruṣa* 'genitive' were most frequent. The *Karmadhāraya* marking the relation of co-referentiality was also frequent. The Sanskrit grammar also provided us certain clues for identifying a compound and its types based on its components.

Finally an important observation from Pāṇini's treatment of Sanskrit grammar was the following. Pāṇini has strived hard to make his grammar as exhaustive as possible by providing rules to handle very rare compounds. So we could take the advantage of both the statistical techniques which perform better with frequent cases and the rule based approach to cover the rare cases.

4 Nominal Compounds in Marathi and Hindi

We used these insights for processing nominal compounds in two major Indian Languages, namely, Hindi and Marathi. Both languages do not have any specific convention for writing the compounds. We find instances of compounds written with components joined together, with a hyphen in between and also with a space in between. Compounds when written as a single word need a segmenter to split it into valid components.

It was observed that a special type of compounds termed as 'Avyayībhāva' are always written as a single word in these languages. Examples of such compounds are *yathāśakti* 'as per the capability', *anurūpa* 'in accordance with'. These are statistically found to be rare in Sanskrit Corpus. We also observe that in Hindi as well as Marathi also such compounds are rare. There could be two ways of accounting for such rare compounds:
1. To have them stored in the lexicon.
2. To have rules with the help of which such compounds can be analysed.
Pāṇini has accounted for such rare compounds with rules which we may use for Hindi and Marathi as well.

In what follows we concentrated only on the compounds written with a space in between. The study was undertaken only to decide the tagset for marking the relation between the components. The tagset for Sanskrit is very exhaustive, covering even the rare compounds. However the purpose of this study was to identify only those tags which are frequent in Hindi and Marathi.

The researchers working on the Cross Lingual Information Retrieval systems among Indian

Languages at IIT Bombay have developed a tool for automatic extraction of Multi Word Expressions from a corpus that uses minimum linguistic tools such as morphological analysers, and POS taggers. The candidates were ranked using Point-wise Mutual Information (PMI) method. Marathi corpus from Tourism domain consisting of 15,925 sentences with 0.325M words was chosen for the experiment. The Multi Word Expression extraction tool gave an initial set of Multi Word Expressions. From these Multi Word Expressions for Marathi, noun compounds were extracted manually, and a study was undertaken to identify the relations between the components. Table 2 lists the identified relations with examples from Marathi.

Dependence relation	Example	Gloss	Meaning
Tādarthya (Purpose)	Praveśa dvāra	Entry door	Door for entry
Karaṇa (Instrument)	Hasta shilpa	Hand Architechture	Architecture made by Hand
adhikaraṇa (Location)	Bhitti chitra	Wall painting	Painting on the wall
samānādhikaraṇa (co-referentiality)	Bauddha dharma	Buddhist religion	Buddhist religion
	sāhasika paryaṭana	adventurous Tourism	adventurous Tourism
	dara varṣī	Every in year	Every year
śeṣa (genitive)	samudra taṭa	Sea bank/shore	Shore of sea
	pāka kalā	Cooking art	Art of cooking
	paryaṭana sthala	tourism place	place of tourism
	upāhāra gṛha	little food house	restaurant

Table 2: Relations in Marathi Noun compounds

As observed in Sanskrit, in Marathi as well we found compounds with genitive case marker, compounds with co-referential components and compounds involving various kinds of dependency relation amongst the components were dominant.

In order to make sure that these relations are sufficient across other ILs, we repeated this study with Hindi. However, this time the compounds were extracted from a Hindi-Urdu dependency treebank being developed at IIIT Hyderabad. Pāṇinian grammar formalism is being followed for the annotation. The treebank has 10,799 sentences consisting of approximately 0.25M words. The compounds in this treebank form a chunk and are annotated with a special label. This made it easy to extract the sentences with compounds. We have examined around 827 sentences and identified 895 noun compounds with two components. Number of unique compounds is 597. Among them 20 are dvandva (copulative) compounds and 15 are cases of reduplication. We observe that compounds can be analysed with genitive (*ṣaṣṭhī sambandha*) for around 45% of times even though we understand that paraphrasing with genitives does not necessarily capture deep semantic relations (see section 3.3 for examples). Nevertheless for the purpose of machine translation genitive paraphrasing may be sufficient because a genitive construct in one language can be mostly translated into a genitive construct in another language whereas a source language compound need not remain a compound in the target language. For example, both the compound *room temperature* and the corresponding genitive construction *temperature of room* can only be translated into *kamare kā tāpamāna* in Hindi. In case of

English-Hindi language pairs, it was observed that in 59% of cases an English Noun compound can be translated into genitive construction in Hindi (Paul et al., 2010). However, for other NLP tasks such as information extraction, question answering etc., genitive relation will not be sufficient and one needs to look for deeper semantic relation. In the present work, we have attempted annotation of deeper semantic relation only when the paraphrase with genitive is illegitimate. The paraphrases were of the following types.

a) with vibhakti (equivalent to post-positions)
b) verb + vibhakti (and not with vibhakti alone)
c) Subtype relation (hyponymy relation)
d) Other kinds of paraphrase

We will discuss each case with suitable examples:

(I) Paraphrasing with vibhakti alone
We come across six classes of vibhaktis[4] which are used for paraphrasing other than the genitive one. Table 3 provides the examples.

(II) Paraphrasing with verb + vibhakti
We find many cases where a meaningful paraphrase is not possible with post-position alone. We have used verbal form along with post-position for meaningful paraphrasing for such cases. For example:

- antarīkṣa yāna vs antarīkṣa meṁ jāne vālā (or ke lie) yāna
  gloss: space ship vs ship that goes into space
- rela saḍaka vs rela calane ke lie saḍaka
  gloss: railway track vs track for the running of train
- rājya sarakāra vs rājya ke lie cunī gayī sarakāra
  gloss: state government vs a government to run the state affairs
- nirvāṇa sthala vs nirvāṇa prāpti (or pāne) ke lie sthala (which according to some annotators can be tagged as nirvāṇa ke lie sthala)
  gloss: nirvāṇa place vs a place where nirvāṇa is attained
- garbha gṛha vs garbha meṁ sthita gṛha
  gloss: inside room vs a room situated inside

(III) Hyponymic Relation
This relation is quite common apart from vibhakti paraphrasing. Hyponymic relations can again be of different nature as exemplified below.

(i) A hyponymic relation which is similar to samānādhikarana (see section 3.3). For example,
a. kāngresa dala vs kāngresa IS_A (nāmaka) dala
gloss: Congress Party
b. taṭarakṣaka bala vs taṭarakṣaka IS_A (nāmka) bala
gloss: Post guard

[4] Vibhakti or post-position can be multi word in Hindi; for example: ke viṣaya meṁ, ke bAre meṁ etc. These multi word expressions are treated as one post-position.

Vibhakti	Meaning	Instances		
		Compound	Gloss	English Translation
se, ke dvārā	With	rimoṭ kanṭrola	Remote Control	Remote Control
		gaisa pīḍita	Gas Victim	Gas Victim
		phoṭo pahacāna	Photo Identity	Identity Card
		dhvani pradūṣaṇa	Noise Pollution	Noise Pollution
		dūrasaṁchāra sevā	Telecom Service	Telecom Service
se	From	sevā nivṛtta	Service Retired	Retired from Service
		karma nivṛtti	Work Retired	Retired from Work
ke lie	For	surakshā bala	Security Forces	Security Forces
		samanvaya samiti	Sensation Committee	Sensation Committee
		ṭurisṭa hāusa	Tourist House	Tourist House
		prajanana aṁga	Reproductive Organs	Reproductive Organs
		śoka saṁvedanā	Mourning Sensation	Condolence
meṁ, para	In, On, At	besa kāmpa	Base Camp	Base Camp
		bāla aparādha	Child Crime	Juvenile Delinquency
ke viṣaya meṁ	About	vidhi śikṣā	Legal Education	Legal Education
		śramika mudde	Labour Issues	Labour Issues
ke sambandha meṁ, sambandhita	About (with regard to)	videśa nīti	Foreign Policy	Foreign Policy
		anusandhāna vibhāga	Research Department	Research Department

Table 3: Paraphrasing Hindi compounds with vibhakti alone

(ii) A hyponymic relation which refers the whole Multi Word Expression as a type of the entity that the head denotes; for example:
a. moṭara boṭa vs moṭara boṭa IS_A (Type of) boṭa (unlike (i), moṭara is not a boṭa)

gloss: Motor Boat
b. prashna cinha vs prashna cinha IS_A (Type of) cinha
gloss:question mark

(IV) Other kinds of paraphrases
We come across some cases where the genitive paraphrase is possible if and only if the modifier can be pluralized. For example,

(a) film abhinetā vs filmoṁ kā abhinetā
gloss: film actor vs actor of film

(b) pujārī samudāya vs pujārīyoṁ kā samudāya
gloss: priest group vs group of priests

(c) cālaka dala vs cālakoṁ kā dala
gloss: driver group vs group of drivers

The heads of (b) and (c) are aggregate nouns and therefore the modifier acquires plural meaning. In case of (a), *film (sg.) kā abhinetā* would mean actor of a particular film; whereas *film abhinetā* as a compound conveys the meaning of 'profession' as in *amitābha eka film abhinetā haiṁ*. The other suitable paraphrase would be *film meṁ kāma karane vālā abhinetā*, where *film* remains singular.

There are institutionalized terms such as *pulisa āyukta, vikāsa saciva* and so on and also borrowed compounds such as *ḍāyala up* 'dial up', *kebal cār* 'cable car', which we have left out of the scope of paraphrasing. Table 4 presents number of occurrence of various paraphrases in the data that we have analysed.

Type	No of Instances	Percentage
Genitive	270	47.12
Paraphrasing with Vibhakti alone	80	13.96
Hyponymic Relation	68	11.86
Paraphrasing with verb + Vibhakti	40	6.98
Copulative	20	3.49
Reduplications	15	2.62
Other kinds of Paraphrases	14	2.44
Difficulty in annotation	66	11.51

Table 4: Analysis of Hindi data for various types of paraphrases

This table further supports our intuition from analysis of Marathi data that the tatpuruṣa 'endocentric with missing case markers' and the copulative compounds are more frequent, providing a very strong empirical support for the development of tagset for semantic annotation of noun compounds. In Hindi we also found a considerable number of compounds which require an additional verb and a post position marker for paraphrasing. It is necessary to study the last category of compounds further in order to enable machines to carry out the analysis and ‘guess’ the missing verb automatically.

5 Conclusion

It is clear from the data analysed for Sanskrit and Hindi that the most dominating type of compounds is the Ṣaṣṭhī Tatpuruṣa (genitive), in both languages. There are cases of Tatpuruṣa which cannot be analysed with genitive and they are paraphrased with various vibhaktis (case markers). We come across quite a number of cases of Karmadhāraya ‘hyponymic’ compounds. It is interesting to note that compounding in Hindi deviates from that in Sanskrit in two aspects. Data analysed for Hindi does not contain any instance of Bahuvrīhi (exo-centric) compound. Second, Hindi data presents many cases where quite a lot of compounds requires a verb as well as vibhakti for its paraphrasing. Such compounds are termed as madhyama-pada-lopī in Sanskrit, and they are found to be rare in Sanskrit. The compounds which were found to be difficult for the annotators should form the part of a lexicon.

References

Butnariu, C. and Veale, T. (2008). A concept-centered approach to noun compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK.

Finin, T. W. (1980). The semantic interpretation of nominal compounds. In *In the Proceedings of the 1st Conference on Artificial Intelligence (AAAI-80)*.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *In the proceedings of the Human Language Technology Conference (HLT)*.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). Classification of semantic relations between nominals. In *Proceedings of The Semantic Evaluation Workshop (SemEval) in Conjunction with ACL*, Prague.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Huet, G. (2009). Formal structure of Sanskrit text: Requirements analysis for a mechanical Sanskrit processor. In Huet, G., Kulkarni, A., and Scharf, P., editors, *Sanskrit Computational Linguistics 1 & 2*. Springer-Verlag LNAI 5402.

Kim, S. N. and Baldwin, T. (2006). Interpreting semantic relation in noun compound via verb semantics. In *Proceedings of ACL/COLING-2006*.

Kulkarni, A. and Kumar, A. (2011). Statistical constituency parser for Sanskrit compounds. In *Proceedings of ICON 2011*. Macmillan Advanced Research Series, Macmillan Publishers India Ltd.

Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A., and Bhattacharyya, P. (2010). Introducing sanskrit wordnet. In Pushpak Bhattacharyya, C. F. and Vossen, P., editors, *Principles, Construction and Application of Multilingual Wordnets, Proceedings of the Global Wordnet Conference, 2010*. Narosa Publishing House, New Delhi.

Kumar, A. (2012). *An automatic Sanskrit Compound Processing*. PhD thesis, University of Hyderabad, Hyderabad.

Kumar, A., Mittal, V., and Kulkarni, A. (2010). Sanskrit compound processor. In Jha, G. N., editor, *Proceedings of the International Sanskrit Computational Linguistics Symposium*. Springer-Verlag LNAI 6465.

Kumar, A., SheebaSudheer, V., and Kulkarni, A. (2009). Sanskrit compound paraphrase generator. In *Proceedings of ICON 2009*.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun compounds*. PhD thesis, Macquarie University, Australia.

Mittal, V. (2010). Automatic sanskrit segmentizer using finite state transducers. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 85–90, Uppsala, Sweden. Association for Computational Linguistics.

Nair, S. and Kulkarni, A. (2010). The knowledge structure in amarakośa. In Jha, G. N., editor, *Proceedings of the International Sanskrit Computational Linguistics Symposium*. Springer-Verlag LNAI 6465.

Nakov, P. (2008). Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceeding of 13th International Conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA-08)*, Varna, Bulgaria.

Nastase, V. and Szpakowicz, S. (2009). The same semantic relations link structurally different realizations of concept. In *Linguistic Issues in Language*.

Paul, S., Mathur, P., and Kishore, S. (2010). Syntactic construct: An aid for translating english nominal compound into hindi. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, Los Angeles, California.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Piscataway.

Séaghdha, D. O. and Copestake, A. (2007). Co-occurrence contexts for noun-compound interpretation. In *Proceedings of the ACL-07 Workshop on a Broader Perspective on Multiword Expression (MWE-07)*, Prague, Czech Republic.

Shastri, G. (2006). *Patañjali's Vyākaraṇa Mahābhāṣya with Kaiyaṭa's Pradīpa and Nāgojibhaṭṭa's Uddyota with the Notes by Guruprasad Shastri (Adhyāya 2)*. Rashtriya Sanskrit Sansthan, New Delhi (reprint of 1938 edition).

Vanderwende, L. (1995). *The Analysis of Noun Sequences Using semantic Information Extracted from on-line Dictionaries*. PhD thesis, Georgetown University.