Clues from Aṣṭādhyāyī for compound type identification

Amba Kulkarni and Anil Kumar

Department of Sanskrit Studies,
University of Hyderabad,
Hyderabad
apksh@uohyd.ernet.in, anil.lalit22@gmail.com

Abstract. Aṣṭādhyāyī has a section of rules which provide conditions for compound formation. These rules are presented from generation point of view. We study these conditions from the point of view of compound type identification. A rule based classifier based on these rules is developed whose performance on some of the compound types is encouraging. These conditions also suggest the type of information lexical databases should contain for automatic language analysis, including a compound classifier.

1 Introduction

Sanskrit is very rich in compound formation. On an average[1] every fifth or sixth word in a Sanskrit sentence is a compound. The compound formation being productive it forms an open-set and as such it is also not possible to list all the compounds in a dictionary. The compound formation involves a mandatory sandhi. But mere sandhi splitting does not help a reader in identifying the meaning of a compound. Typically a compound does not code the relation between its components explicitly. To understand the meaning of a compound, it is necessary to identify its components and discover the relation between them.

A worth noting contribution in the field of Sanskrit compounds is by the Department of Indology of French Institute, Pondichery. It has developed a searchable electronic version of Pāṇinīya Udāharaṇakośa(Grimal, 2008) providing generation of approximately 4,400 compounds from Mahābhāṣya, Kāśikā and Kaumudī. This database is very much useful in understanding the process of compound formation. Gillon(2009) suggests tagging of compounds by enriching the context free rules. The idea here is to mark the missing case marker and the head, and specify the enriched category of the compound.

[1] This estimate is based on the manually tagged corpus available with Sanskrit Consortium.

The Sanskrit Consortium under the Government of India sponsored project has developed a tagset for Sanskrit compounds (Ramakrishnamacharyulu et al. 2012). This tagset has 56 tags (Appendix A), which are sub-classifications of the major 4 types *Avyayībhāva, Tatpuruṣa, Bahuvrīhi* and *Dvandva.* The sub-classification is guided by the paraphrase of compounds. For example, the paraphrase in *Tatpuruṣa* is a function of the vibhakti. Thus there are 7 sub-types under the major type *Tatpuruṣa.*

The question we are addressing in this paper is, is it possible to decide the type / sub-type of a compound in the form of string of phonemes without any accents, merely on the basis of its components? The famous example of *Rāmeśvaraḥ* as illustrated below tells us that mere components do not help us in deciding the type of a compound. *Rāmeśvaraḥ*, depending on the context may mean

a) Rāmaḥ ca asau īśvaraḥ ca,
b) Rāmaḥ īśvaraḥ yasya saḥ, or
c) Rāmasya īśvaraḥ.

In the first case it is *Karmadhāraya*, in the second case it is *Bahuvrīhi* and in the third case it is *Ṣaṣṭhī-Tatpuruṣha*! Though the components are the same in all the three cases, relation between these components is determined from the wider context or possibly through an accent.

It is also important to note the level of semantics the compound types deal with. Consider the compounds *rājapuruṣaḥ, Daśarathaputraḥ*, and *vṛkṣaśākhā.* In the first case the relation between *rājan* and *puruṣa* is that of servant-master (*sevya-sevaka*), in the second the relation between *Daśaratha* and *putraḥ* is of father-son (*pitā-putra*) and in the third case the relation between *vṛkṣa* and *śākhā* 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, Daśarathasya putraḥ* and *vṛkṣasya śākhā*, and thus these are classified as *Ṣaṣṭhī-Tatpuruṣa.* 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 can 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 given in Appendix A, the question is, to what extent is it possible to decide the relation between the words only on the basis of components involved? Anil et al. (2010) have reported that the components do provide an information that is useful for classifying the compounds. For classification, a manually tagged corpus was used as a training data. The results were encouraging, but it was observed that use of rule based classifier along with the statistical data should provide better results. In this paper we study the relevant *sūtra*s from Aṣṭādhyāyī for clues.

2 Clues from Pāṇini's Aṣṭādhyāyī

Pāṇini describes the process of compound formation starting from the paraphrase of a compound. The process starts with an *alaukika vigraha*. It involves deciding the type, the order of the components, the elision of the vibhakti of intermediate components, assignment of *svaras*, addition of certain suffixes to change the gender if necessary.

The *sūtra*s dealing with compounds then can be broadly classified into two types

a) *sūtra*s providing semantic conditions for compound formation, and
b) *sūtra*s dealing with the process of compound formation involving
 (i) decision of word order,
 (ii) *vibhakti* deletion,
 (iii) *svara* assignment,
 (iv) changes in gender and number.

The second type of *sūtra*s are useful from generation point of view. They deal with the morphology and phonology. A close look at the Pāṇini's *sūtra*s of first type provides us a lot of semantic clues. For example, look at the following *sūtra*s.

dvitīyā śritātītapatitagatātyastaprāptāpannaiḥ(2.1.24)
tṛtīyā tatkṛtārthena guṇavacanena(2.1.30)
annena vyañjanam(2.1.33)
pañcamī bhayena(2.1.36)

Each of these *sūtra*s gives a criterion for formation of a compound of a particular type either in terms of semantics of the components involved or as a list of words as the first or the second component. Applied reversely, given the components of a compound, these semantic conditions / list provide a clue for deciding the possibility of a compound to be of a particular type. To make the point clear, *pañcamī bhayena*(2.1.36) states that a word ending in fifth case may optionally get compounded with a word indicating fear. For example, *corāt bhayam* is compounded optionally as *corabhayam*. Now, when we analyse this compound, based on the fact that *bhaya* is the second component, we guess that this compound might have been formed by *pañcamī bhayena*(2.1.36), and a Pāṇinian scholar would typically verify that this is so. Verification rules out false positives. In the absence of a compound generator, we skip the verification part and do only guess. We examine below the relevant *sūtra*s from the Aṣṭādhyāyī to decide whether the clues they provide are sufficient enough for decision making. It is necessary to keep one deviation from Pāṇini in mind. When a compound W is split as w_1-w_2, the morphological features associated with W are accessible through the analysis of w_2. So when we specify the conditions, the gender and number information corresponding to w_2 actually corresponds to the compound W. Most of the *sūtra*s specify a pattern or a condition on either or both the *pūrvapada* as well as *uttarapada*. The *sūtra*s we discuss are from *avyayībhāvaḥ*(2.1.5) to *cārthe dvandvaḥ*(2.2.29). Of these, *sūtra*s which do not employ any condition but serve only as an *adhikāra sūtra* such as *avyayībhāvaḥ*(2.1.5) or *tatpuruṣaḥ*(2.1.21), etc. will not be discussed.

2.1 Avyayībhāva

The sūtras from *avyayaṃ vibhakti-samīpa-samṛddhi-vyṛddhyarthābhāvātyaya-asaṃprati-śabdaprādurbhāva-paścādyathānupurvya-yaugapadya-sādṛśya-sampatti-sākalya-antavacaneṣu*(2.1.6) to *anyapadārthe ca saṃjñāyām*(2.1.21) provide conditions for the formation of an *Avyayībhāva* compound. Table 1 summarizes the conditions on the *samāsa pūrvapada* and *uttarapada* described in these rules. Due to *avyayībhāvaśca*(2.4.18), every *Avyayībhāva* compound is in neuter gender. During analysis process, since we split W as W_1-W_2, W_2 will be in neuter gender. Thus a necessary condition for any compound to be an *Avyayībhāva* is that W_2 should be in neuter gender.

Sūtra tiṣṭhadguprabhṛtīni ca(2.1.17) provides an exceptional list of *Avyayībhāva* compounds. In order to use the sūtras *saṃkhyā vaṃśyena*(2.1.19) and *nadībhiśca*(2.1.20) for deciding the type of a compound, one needs a list of family names and a list of names of rivers.

Sūtra *yathā'sādṛśye*(2.1.7) deals with a broader context. The word *yathā* has four different senses viz. i) *yogyatā* 'ability', ii) *vīpsā*, iii) *padārthānativṛtti*, and iv) *sādṛṣya* 'similarity'. *yathā'sādṛśye*(2.1.7) states that *yathā* is invariably compounded with a case inflected word in a meaning other than *sādṛsya* 'similarity' as in *yathāvṛddham* 'every old person'. Since *yathā* in other meanings also gets compounded, as in *yathāśakti*, given a compound with *yathā* as the first component, it is not possible to decide the meaning of *yathā* only on the basis of the following component, without looking at the complete context. Hence it is not possible to decide the meaning of this compound. Nevertheless we can always mark this as an *Avyayībhāva*.

Anyapadārthe ca saṃjñāyām(2.1.21) puts a condition that a river name may get compounded with another noun referring to *anyapadārtha* 'a totally new thing', and one needs a broader context to decide whether such a compound is an *Avyayībhāva* or not.

S.No.	**Sūtra Number**	**Conditions**		**Type**
		First Component	**Second Component**	
1	2.1.6[2]	Any of the prefixes or indeclinables such as pra, parā, apa, yathā etc.	Neuter gender in first case	A1
2	2.1.7[3]	yathā	Neuter gender in first case	A1
3	2.1.8[4]	yāvat	Neuter gender in first case	A1
4	2.1.9[5]	-	prati	A2
5	2.1.10[6]	akṣa, Śalākā or a numeral such as eka, dvi etc.	pari	A2
6	2.1.12[7]	apa, pari, bahis or word ending with añc	Neuter gender in fifth-case	A1
7	2.1.13[8]	āṅ	Neuter gender in fifth-case	A1
8	2.1.14[9]	abhi, prati	Neuter gender in first case	A1
9	2.1.15[10]	anu	Neuter gender in first case	A1
10	2.1.16[11]	anu	Neuter gender in first case	A1
11	2.1.17[12]	list of tiṣṭhadguprabhṛtī		A3
12	2.1.18[13]	pāre, madhye	Neuter gender in 5th-case	A7

[2] *avyayaṃ vibhakti-samīpa-samṛddhi-vyṛddhyarthābhāvātyayāsamprati-śabdaprādurbhāva-paścādyathā-ānupụrvya-yaugapadya-sādṛśya-sampatti-sākalya-antavacaneṣu*

[3] *yathā'sādṛśye*

[4] *yāvadavadhāraṇe*

[5] *suppratinā mātrārthe*

[6] *akṣaśalākāsaṃkhyāḥ pariṇā*

[7] *apaparibahirañcavaḥ pañcamyā*

[8] *āṅ maryādā'bhividhyoḥ*

[9] *lakṣaṇenā'bhipratī ābhimukhye*

[10] *anuryatsamayā*

[11] *yasya cāyāmaḥ*

[12] *tiṣṭhadguprabhṛtīni ca*

[13] *pāre madhye ṣaṣṭhyā vā*

12.1	2.1.18[14]	-	pārāt, madhyāt	A7
13	2.1.19[15]	Numerals such as eka, dvi	family name	A6
14	2.1.20[16]	Numerals such as eka, dvi	name of a river	A6
15	2.4.83[17]	Any of the prefix or indeclinables such as pra, parā, etc.	prātipadika ending in ’a’ in 3rd, 5th, 7th	A7

Table 1: Conditions for the compound *Avyayībhāva*

2.2 Tatpuruṣaḥ

The sūtras from *dvitīyā śritātītapatitagatātyastaprāptāpannaiḥ*(2.1.24) to *ktvā ca*(2.2.22) provide conditions for the formation of *Tatpuruṣa-samāsa*. We treat *Karmadhāraya*, which is a special case of *Tatpuruṣa* compound, separately. Hence, *sūtra*s from *pūrvakālaikasarvajaratpurāṇanavakevalāḥ samānādhikaraṇena*(2.1.49) to *pūrvāparādharottaramekadeśinaikādhikaraṇe*(2.2.1) are dealt with in the next section. Majority of *sūtra*s have a requirement of a *kṛdanta* ending in *kta* suffix (a past participle form of a verb). Out of the 38 sūtras dealing with *Tatpuruṣa*, only 28 can be used to decide the compound type automatically. The remaining 10 sūtras require extra-linguistic information. For example, the sūtra *atyantasaṃyoge ca*(2.1.29) states, if there is an invariable uninterrupted connection (*atyantasaṃyoga*) between the meaning of the two components, then the resulting compound is of type *Tatpuruṣa*. It is not possible to decide whether the connection is invariable and uninterrupted or not, only on the basis of its components. On similar reasoning to classify a compound *śaṅkulākhaṇdaḥ* ‘cut by knife’ or *dhānyārthaḥ* ‘wealth acquired by grain’ on the basis of *tṛitīyātatkṛtārthena guṇavacanena*(2.1.30) as a *Tṛtīyā Tatpuruṣa*, one needs an information that *śaṅkulā* is an instrument used for cutting, or wealth can be acquired by grains. Few other sūtras that require semantic information are :-

(a) *kartṛkaraṇe kṛtā bahulam*(2.1.32)
(b) *kṛtyairadhikārthavacane*(2.1.33)
(c) *annena vyañjanam*(2.1.34)
(d) *bhakṣyeṇa miśrikaraṇam*(2.1.35)
(e) *kṛtyairṛṇe*(2.1.43)
(f) *saṃjñayām*(2.1.44)
(g) *kṣepe*(2.1.47)

[14] *pāre madhye ṣaṣṭhyā vā*

[15] *saṃkhyā vaṃśyena*

[16] *nadībhiśca*

[17] *nāvyayībhāvādato’mtvapañcamyāḥ*

Sūtra nañ(2.2.6) for *Nañ Tatpuruṣa* poses a special problem with respect to splitting. While the initial *an* may provide an useful clue for the split, initial *a* may lead to over-generation during sandhi splitting. Once split, of course, it is easy to mark such compounds as *Nañ Tatpuruṣaḥ*.

*Sūtra*s from *Ṣaṣṭhī*(2.2.8) to *nityaṃ krīḍājīvikayoḥ*(2.2.17) deal with the *Ṣaṣṭhī Tatpuruṣa* compounds. Among these only one *sūtra yājakādibhiśca*(2.2.9) is assertive providing a condition for the formation of a compound. *Nityaṃ krīḍājīvikayoḥ*(2.2.17) requires a specific semantics that the resulting compound should indicate a game or a means for livelihood. If the lexicon provides explicitly this information, then only automatic detection of such a compound as a *Ṣaṣṭhī Tatpuruṣaḥ* is possible. All other *sūtra*s prohibit the formation of this compound.

Kugatiprādayaḥ(2.2.18) provides conditions on the first component that it can be either a *ku*, or a word which may be termed as *gati* or a list of indeclinables *pra* etc..

Upapadamatiṅ(2.2.19) states a condition in terms of an *upapada*. An *upapada* is a word referring to the word in seventh case in the *sūtra*s that prescribe a *kṛdanta* suffix[18]. Thus this condition refers to an internal stage during the derivational process. Then, given a compound - a generated word, how can we decide whether its *pūrvapada* is an upapada or not. It is not possible to decide whether the *pūrvapada* is an upapada or not, unless we look at the involved process. If *uttarapada* is analysed with these *kṛdanta* suffixes, the derivation of which involves a notion of *upapada*, we may guess the compound to be of type *Upapada Tatpuruṣa*. Further, since these *kṛt* suffixes produce words that are bound morphemes, a morphological analyser handling these bound morphemes should help in deciding whether the given compound is an *Upapada Tatpuruṣa* or not.

Next *sūtra amaivāvyayena*(2.2.20) also puts a condition on *Upapada-tatpuruṣa* compounds, where the requirement is that the *uttarapada* is an indeclinable derived from an *am* ending *kṛt* suffix. *Tṛtīyāprabhṛtīnyanyatarasyām*(2.2.21) extends this condition on *pūrvapada* to other *upapadas* due to the *sūtra*s *upadaśastṛtīyāyām*(3.4.47) - *anvacyānulomye*(3.4.64) as well. *Ktvā ca* (2.2.22) further extends it to a *kṛt* suffix *ktvā*. As stated earlier, it is not possible to decide whether the *pūrvapada* is an *upapada* or not. But in this particular case, we have another clue. Due to *samāse anañpūrve ktvo lyap*(7.1.37), the second component is in *lyap*, and its *pūrvapada* is not a prefix. This provides a clue for guessing the *Upapada Tatpuruṣa* correctly. *Alaṁkṛtya* is an exception.

Table 2 lists the *sūtra*s where it is possible to identify the compound type as *Tatpuruṣa* based on the components and their morphological features mentioned in the *sūtra*s. The *sūtra pātresamitādayaśca*(2.1.48)

[18] *tatra-upapadam saptamīsthānam* 3.1.92

provides a list of exceptional *Tatpuruṣa* compounds. In *caturthī tadarthārthabalihitasukharakṣitaiḥ*(2.1.36), to decide whether the relation between the two components is of *tadartha*[19] or not is difficult. The other conditions of *caturthī tadarthārthabalihitasukharakṣitaiḥ*(2.1.36) which lists various words for the possible candidate are easy to implement mechanically. For automatic detection of a compound type, as is evident from the following conditions, we also require a list of words denoting time, time indicating words which can be used as a measure, parts of a day and night, list of names of river, list of family names, words denoting numbers, etc.

S.No.	**Sūtra Number**	**Conditions**		**Type**
		First Component	**Second Component**	
1	2.1.24[20]	-	śrita, atīta, patita, gata, atysta, prāpta, āpanna	T2
2	2.1.25[21]	svayam or svāyam	kṛdanta with kta pratyaya	T2
3	2.1.26[22]	khaṭvā	kṛdanta with kta pratyaya	T2
4	2.1.27[23]	sāmi	kṛdanta with kta pratyaya	T2
5	2.1.28[24]	a word denoting time	kṛdanta with kta pratyaya	T2
6	2.1.31[25]	-	pūrva, sadṛśa, sama, ūnārtha, kalaha, nipuṇa, miśra, ślakṣṇa	T3
7	2.1.36[26]	-	artha, bali, hita, sukha, rakṣita	T4
8	2.1.37[27]	-	bhaya, bhīti, bhī	T5
9	2.1.38[28]	-	apeta, apoḍha, mukta, patita, apatra	T5

[19] *tadarthena prakṛtivikṛtibhāva samāsaḥ ayam iṣyate.* Kāśikā(2.1.36)

[20] *dvitīyā śritātītapatitagatātyastaprāptāpannaiḥ*

[21] *svayaṃ ktena*

[22] *khaṭvā kṣepe*

[23] *sāmi*

[24] *kālāḥ*

[25] *pūrvasadṛśasamonārthakalahanipuṇamiśraślakṣṇaiḥ*

[26] *caturthī tadarthārthabalihitasukharakṣitaiḥ*

[27] *pañcamī bhayena*

[28] *apetāpoḍhamuktapatitāpatrastairalpaśaḥ*

10	2.1.39[29]	stoka, antika, dūrārtha, kṛcchra	krdanta with kta pratyaya	T5
11	2.1.40[30]	-	from the list of śauṇḍa gaṇa	T7
12	2.1.41[31]	-	siddha, śuṣka, pakva, bandha	T7
13	2.1.42[32]	-	dhvāṅkṣa	T7
14	2.1.45[33]	words denoting part of a day or night	kṛdanta with kta pratyaya	T7
15	2.1.46[34]	tatra	kṛdanta with kta pratyaya	T7
16	2.1.48[35]	list of exceptional compounds pātresamitā...		T7
17	2.2.2[36]	ardha	-	T1
18	2.2.6[37]	a, an	-	Tn
19	2.2.7[38]	īṣat	-	T
20	2.2.9[39]	-	list of words from yājaka gaṇa	T6
21	2.2.18[40]	ku, kā, pra etc., may be termed as *gati*	list of words from yājaka gaṇa	T6
22	2-2.19[41]	-	a bound morpheme with special kṛt suffix(es)	U
23	2.2.20[42]	word ends with Ṇamul and khamuṅ suffix	a kṛt suffix ending in am	U
24	2.2.21[43]	word ends with third case	a kṛt suffix ending in am	U

[29] *stokāntikadūrārthakṛcchrāṇi ktena*

[30] *saptamī śōṇḍaiḥ*

[31] *siddhaśuṣkapakvabandhaiśca*

[32] *dhvāṅkṣeṇa kṣepe*

[33] *ktenāhorātrāvayavāḥ*

[34] *tatra*

[35] *pātresamitādayaśca*

[36] *ardhaṃ napuṃsakam*

[37] *nañ*

[38] *īṣadakṛtā*

[39] *yājakādibhiśca*

[40] *kugatiprādayaḥ*

[41] *upapadamatiṅ*

[42] *amaivāvyayena*

[43] *tṛtīyāprabhṛtīnyanyatarasyām*

25	2.2.22[44]	word ends with third case	word ends with kṛt suffix kṛtvā or lyap	U

Table 2: Conditions for the compound *Tatpuruṣa*

2.3 Karmadhāraya

Twenty seven *sūtra*s from *pūrvakālaikasarvajaratpurāṇanavakevalāḥ samānādhikaraṇena*(2.1.49) to *mayūravyaṃsakādayaśca*(2.1.72) and *pūrvāparādharottaramekadeśinaikādhikaraṇe*(2.2.1), except *saṃkhyāpūrvo dviguḥ*(2.1.52) indicating a *dvigu*, provide conditions for marking a compound as a *Karmadhāraya*. Out of these 23, 15 *sūtra*s provide conditions which involve a controlled semantics that can be handled by a computer. Other 8 *sūtra*s either deal with open lists, or involve deeper semantics. For example, words which share common properties and thus can be compared, or compounds with abusing words as one of the components, or words indicating a part-whole relation form a *Karmadhāraya* compound. Sanskrit WordNet(Kulkarni et al. 2010) and also Amarakośa which is tagged for semantic information(Sivaja et al. 2010) are some of the resources where one can get the pairs with part-whole relation. With the availability of this information, *pūrvāparādharottaramekadeśinaikādhikaraṇe*(2.2.1) may be implemented. However, to get an information about the words sharing common properties is not within reach with the available e-lexicon. Table 3 shows various conditions on the components of *Karmadhāraya* compounds as stated in the above range of *sūtra*s. As we note, *poṭāyuvatistokakatipayagṛṣṭidhenuvaśāvehadbaṣkayaṇīpravaktṛśrotriya-adhyāpaka-dhūrtairjātiḥ*(2.1.65) requires the *pūrvapada* to be a word indicating a class. *Praśaṃsāvacanaiśca*(2.1.66) puts a semantic constraint on the second component demanding it to be a praising word. Similarly *varṇo varṇena*(2.1.69) requires both the components from a list of color words, while *catuṣpādo garbhiṇyā*(2.1.71) needs the first component to be a four-legged animal. So only if the lexicon is rich with this information, one can use these *sūtra*s for compound type identification. *Mayūravyaṃsakādayaśca*(2.1.72) provides an enumeration of special compounds which otherwise do not have any common semantic condition.

[44] *ktvā ca*

S.No.	Sūtra	Conditions		Type
	Number	First Component	Second Component	
1	2.1.49[45]	pūrva, kāla, eka, sarva, jarat, purāṇa, nava, kevala	-	K1
2	2.1.50[46]	Words indicating a number or direction	-	K3
3	2.1.56[47]	-	a word from vyāghra gaṇa	K5
4	2.1.58[48]	pūrva, apara, prathama, carama, jaghanya, samāna, madhya, madhyama, vīra	-	K1
5	2.1.59[49]	a word from śreṇyādi gaṇa	a word from kṛt gaṇa	K
6	2.1.60[50]	a word ending in kta pratyaya	negation of the first component	K2
7	2.1.61[51]	sat, mahat, parama, uttama, utkṛṣṭa	-	K1
8	2.1.62[52]	-	vṛndāraka, nāga, kuñjara	K2
9	2.1.63[53]	katara, katama	-	K1
10	2.1.64[54]	kim	-	K1
11	2.1.65[55]	a class denoting word	poṭā, yuvati, stoka, katipaya, gṛṣṭi, dhenu, vaśā, vehat, baṣkayaṇī, pravaktṛ, śrotriya, adhyāpaka, dhūrta	K2

[45] *pūrvakālaikasarvajaratpurāṇanavakevalāḥ samānādhikaraṇena*

[46] *diksaṃkhye saṃjñāyām*

[47] *upamitaṃ vyāghrādibhiḥ sāmānyāprayoge*

[48] *pūrvāparaprathamacaramajaghanyasamānamadhyamadhyamavīrāśca*

[49] *śreṇyādayaḥ kṛtādibhiḥ*

[50] *ktena nañviśiṣṭenānañ*

[51] *sanmahatparamottamotkṛṣṭāḥ pūjyamānaiḥ*

[52] *vṛndārakanāgakuñjaraiḥ pūjyamānam*

[53] *katarakatamō jātiparipraśne*

[54] *kiṃ kṣepe*

[55] *potāyuvatistokakatipayagṛṣṭidhenuvaśāvehadbaṣkayaṇīpravaktṛśrotriya-adhyāpaka-dhūrtairjātiḥ*

12	2.1.66[56]	a class indicating word	words signifying praise such as matallikā, macarcikā, uddha, tallaja	K2
13	2.1.67[57]	yuvan	khalati, palita, valina, jarati	K2
14	2.1.69[58]	a color denoting word	a color denoting word	K3
15	2.1.70[59]	kumāra	a word from śramaṇa gaṇa	K2
16	2.1.71[60]	name of a 4-legged animal	garbhiṇī	K2
17	2.1.72[61]	Exceptional compounds belonging to mayūravyaṃsaka list		K
18	2.2.1[62]	pūrva, para, adhara, uttara	a word indicating an object having parts	K
19	2.2.3[63]	pūraṇa saṁkhyā	-	K
20	2.2.4[64]	prāpta, āpanna	jīvikā	K
21	2.2.4[65]	jīvikā	prāpta, āpanna	K
22	2.2.5[66]	words indicating time used as a measure	words used for measuring	K

Table 3: Conditions for the compound *Karmadhāraya*

2.4 Bahuvrīhi

Among the six *sūtra*s from *śeṣo bahuvrīhiḥ*(2.2.23) to *tena saheti tulyayoge*(2.2.28), dealing with *Bahuvrīhi*, *śeṣo bahuvrīhiḥ*(2.2.23) is an *adhikāra sūtra* and *anekamanyapadārthe*(2.2.24) lays down the general condition for the formation of *Bahuvrīhi.* Conditions stated in *saṃkhyayāvyayāsannādūrādhikasaṃkhyāḥ saṃkhyeye*(2.2.25) and

[56] *praśaṃsāvacanaiśca*

[57] *yuvā khalatipalitavalinajaratībhiḥ*

[58] *varṇo varṇena*

[59] *kumāraḥ śramaṇādibhiḥ*

[60] *catuṣpādo garbhiṇyā*

[61] *mayūravyaṃsakādayaśca*

[62] *pūrvāparādharottaramekadeśinaikādhikaraṇe*

[63] *dvitīyatṛtīyacaturthaturyāṇyanyatarasyām*

[64] *prāptāpanne ca dvitīyayā*

[65] *prāptāpanne ca dvitīyayā*

[66] *kālāḥ parimāṇinā*

diṅnāmānyantarāle(2.2.26) may be used for the detection of *Bahuvrīhi*. *tatra tenedamiti sarūpe*(2.2.27) states that two similar/homophonous words in locative or instrumental case are compounded in the sense of 'this happens with it/there'. eg *muṣṭīmuṣṭi, keśākeśi*, etc. Such compounds are very rare and hence may be treated as exceptions. Table 4 gives a list of possible conditions for detecting a few type of *Bahuvrīhi* compounds.

S.No.	**Sūtra**	**Conditions**		**Type**
	Number	**First Component**	**Second Component**	
1	2.2.25[67]	saṁkhyeya	indeclinable, āsanna, adūra, adhika, saṃkhyā	Bvs
2	2.2.26[68]	name of a direction	name of a direction	Bsd
3	2.2.28[69]	sa	-	BvS

Table 4: Conditions for the compound *Bahuvrīhi*

3 Observations

After going through the relevant *sūtra*s, we observe that the conditions stated by Pāṇini fall under the following categories.

1. A restricted list of words 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 from the point of view of e-lexicon building. The fourth category provides us clues for the important types of relations. Efforts such as Sanskrit Wordnet or on marking semantic information in various kośas such as Amarakośa are concerned about lexical as well as semantic relations. In the *sūtra*s related to compounds, we found the mention of following semantic relations.

(i) *viśeṣaṇa-viśeṣya-bhāva*
(ii) *upamāna-upameya-bhāva*

[67] *saṃkhyayāvyayāsannadūrādhikasaṃkhyāḥ saṃkhyeye*
[68] *diṅnāmānyantarāle*
[69] *tena saheti tulyayoge*

(iii) *avayava-avayavī-bhāva*
(iv) instrument-action relation

The fifth category of conditions mentions certain semantic properties such as

(a) a number
(b) name of a river
(c) a family name
(d) a direction
(e) an abusing word
(f) a praising word
(g) a 4-legged animal
(h) a color word
(i) a class (*jāti*)
(j) an adjective

4 Evaluation

We implemented a rule based classifier based on the above conditions. Manually tagged corpus of size 600K is available with the Sanskrit Consortium. It contains around 64K compounds of two components tagged in context using the tagset specified in Appendix A. Table 5 gives the distribution of high frequent 5 tags.

Tag	**# of words**	**% of words**
T6	26,097	40.56
K1	7,909	12.29
Bs6	4,113	6.39
Tn	3,801	5.9
U	2,782	4.32

Table 5: Distribution of Fine-grain-Tags

We observe that out of the 64K compounds, around 30K compounds were either of type T6 or Bs6. Pāṇini's sūtras covering these compounds are too general[70]. It is the remaining type of compounds for which Pāṇini's sūtras provide semantic criterion; and it has been reported in Anil et al. (2010) that the performance of statistical parsers on these type of compounds is very poor due to scarcity of the data. We test these compounds for rule based tagger. We tested our classifier for 11 tags, and their performance is reported in Table 6. The poor performance corresponding to *Tatpuruṣa* compounds is due to non-availability of semantic information. For example, *Tṛtīyā-Tatpuruṣa* compound needed an information that the first component be a possible instrument for an action denoted by a verb in the second component. In the absence of such information, these compounds can not be classified.

[70] *ṣaṣṭhī* (2.2.8) and *anekamanyapadārthe* (2.2.24)

Sr. No.	Tag	Manually tagged	Tagged by m/c	Correct instances	Precision	Recall	F-measure
1	Tn	3,801	3,770	3,760	99.73	98.92	99.32
2	Tk	47	47	46	97.87	97.87	97.87
3	A1	1,004	1,045	954	91.29	95.01	93.11
4	T4	1,033	1,033	902	87.31	87.31	87.31
5	K1	7,436	2,733	2,365	86.53	31.80	46.50
6	T5	335	100	48	48.00	14.32	22.05
7	T1	132	67	29	43.28	21.96	29.13
8	T2	264	106	37	34.90	14.01	19.99
9	T3	2,334	251	65	25.89	2.78	5.02
10	T7	1,234	31	7	22.58	0.56	1.09
11	A7	13	70	13	18.57	100.00	31.32
		17,633	**9,253**	**8,226**	**88.9**	**46.65**	**61.19**

Table 6: Performance of rule based tagger on less frequent compounds

5 Conclusion

Classification of compounds is a major task in deciding the meaning of a compound. Anil et al. (2010) reported that the precision of a statistical classifier with 55 tags is 63.0%, if we consider only the first rank. If we allow the first three ranks, the performance goes up to 81.1%. Statistical taggers perform well provided the training data is sufficient. So their performance goes down on compounds of rare type. On the other hand, a rule based classifier can perform well on rare type of compounds as well. As we note above, the performance of the rule based classifier is reasonably good for the tags which are less frequent. If a semantically tagged lexicon is made available, the performance of this classifier may increase further. And a proper combination of the two methods should produce better results.

6 Acknowledgment

We thank the anonymous reviewers for their suggestions and feedback. This work is a part of the Sanskrit Consortium project entitled ‘Development of Sanskrit computational tools and Sanskrit-Hindi Machine Translation system’ sponsored by TDIL Programme, DIT, Government of India.

References

1. Bhat, G.M., 2006 : *Samāsaḥ*. Samskrita Bharati, Bangalore, Karnataka.
2. Grimal, F., Sarma, V. V. and Lakshminarasimham, S., 2008: *Pāṇiniyavyākaraṇodāharaṇakoṣaḥ* (Vol. 2: Samāsaprakaraṇam. The book of compound words), French Institute of Pondichery, Pondichery.

3. Gillon, B.S., 2007:*Exocentric Compounds in Classical Sanskrit.* In: Proceeding of the First International Symposium on Sanskrit Computational Linguistics(SCLS-2007), Paris, France.
4. Gillon, B.S., 2009: *Tagging Classical Sanskrit Compounds.* In: Sanskrit Computational Linguistics 3, pages 98-105, Springer-Verlag LNAI 5406.
5. Jha, B.G., 1990 : *Samāsa-sandarśikā.* Chowkhamba Surabharati Prakashan, Varanasi, UP.
6. Joshi, S.D. and Roodbergen, J.A.F., 1969 : *The Vaiyākaraa-mahābhāṣya (avyayībhāvatatpuruṣāhnika).* University of Poona, Poona, Maharastra.
7. Joshi, S.D. and Roodbergen, J.A.F., 1973 : *The Vaiyākaraa-mahābhāṣya (Tatpuruṣāhnika).* University of Poona, Poona, Maharastra.
8. Joshi, S.D., Roodbergen, J.A.F., 1996 : *The Aṣṭādhyāyī of Pāṇini - Volume V and VI.* Sahitya Academy, New-Delhi (India).
9. Kulkarni, A. P., Kumar, A., Sheeba, V. 2009: *Sanskrit compound paraphrase generator.* In: Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India.
10. Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A. and Bhattacharya, P., 2010: *Introducing Sanskrit WordNet.* In: Principles, Construction and Application of Multilingual Wordnets: Proceedings of the 5th Global Wordnet Conference, Narosa Publishers, India. ISBN : 978-81-8487-083-1
11. Kumar, A., Mittal, V., Kulkarni, A. P., 2010: *Sanskrit Compound Processor.* In : Proceedings of 4i-SCLS 2010: 4^{th} International Sanskrit Computational Linguistics Symposium, Springer-Verlag LNAI 6465.
12. Mahavira, June 1978 : *Pāṇini as Grammarian (With special reference to compound formation).* Bharatiya Vidya Prakashan [Delhi - Varanasi], India.
13. Mimamsaka, Yudhisthir : *Mahābhāṣyam (with Hindi commentory) - I, II and III parts.* Ramlal Kapur Trust, Sonepat, Haryana.
14. Murty, M.S., 1974 : *Sanskrit Compounds-A Philosophical Study.* Chowkhamba Sanskrit Series Office, Varanasi(India).
15. Pande, G.D. : *Aṣṭādhyāyī of Pāṇini.* Chowkhamba Surabharati Prakashan, Varanasi, UP.
16. Pandit Ishvarachandra, 2004 : *Aṣṭādhyāyī.* Chaukhamba Sanskrit Pratisthan, Delhi.
17. Ramakrishnamacharyulu, K. V., Kulkarni, A. P., Kulkarni, T., Kumar, A.:*Guidelines for Tagging of Sanskrit Compounds prepaired for Sanskrit Consortium dated 12.03.2012* (Unpublished).
18. Sharma, Vasudeva L.S.P., 1908: *The Siddhānta Kaumudī (With Tatvabodini commentory).* Tukaram Javaji, Proprietor of Javaji Dadaji's "Nirṇayasāgar" Press, Bombay.
19. Shastri, Pt. Guru Prasad, 2006 : *Vyākaraa-mahābhāṣyam (Only Samāsaprakaraṇam).* Rashtriya Sanskrit Sansthan, New-Delhi.
20. Nair, S.S. and Kulkarni, A.P., 2010:*The Knowledge Structure in Amarakośa.* In Proceedings of the International Sanskrit Computational Linguistics Symposium, Springer.
21. Tarkavachaspati, Taranatha, 1812-1885 : *Vācaspatyam.* Chowkhamba Sanskrit Series, Varanasi, UP.
22. Tripathi, V. P., 1991 : *Samāsa-vṛtti-vimarśaḥ.* Sampurnananda Sanskrit Vishvavidyalaya, Varanasi, UP.
23. Vasu, S. C., 1891 : *The Aṣṭādhyāyī of Pāṇini (Translated into English).* Indian Press, Allahabad, UP.

24. Vasu, S. C., : *The Siddhānta Kaumudī of Bhattoji Dikṣita*. Motilal Banarsidas Publishers, New Delhi.

A Compound Tagset

Semantic classifications of Sanskrit compounds					
Avyayībhāva			**Tatpuruṣa**		
1	avyaya-pūrvapada	A1	1	prathamā	T1
2	avyaya-uttarapada	A2	2	dvitīyā	T2
3	tiṣṭhadguprabhṛti	A3	3	tṛtīyā	T3
4	saṃkhyāpūrvapada-nadyuttarapada	A4	4	caturthī	T4
5	nadyuttarapada-anyapadārthasaṃjñāyām	A5	5	pañcamī	T5
6	saṃkhyāpūrvapada-vaṃśyottarapada	A6	6	ṣaṣṭhī	T6
7	pāre-madhye-pūrvapadaṣaṣṭhyuttarapada	A7	7	saptamī	T7
Bahuvrīhi			8	nañ	Tn
1	dvitīyārtha	Bs2	9	prādi	Tp
2	tṛtīyārtha	Bs3	10	ku	Tk
3	caturthyarthabahuvrīhi	Bs4	11	gati	Tg
4	pañcamyartha	Bs5	12	taddhitārthadvigu	Td
5	ṣaṣṭhyartha	Bs6	13	uttarapadadvigu	Tdu
6	saptamyartha	Bs7	14	samāhāradvigu	Tds
7	digvācaka	Bsd	15	upapada	U
8	saṃkhyobhayapada	Bss	16	dvitīyopapada	U2
9	upamānapūrvapada	Bsu	17	tṛtīyopapada	U3
10	praharaṇaviṣayaka	Bsp	18	caturthyopapada	U4
11	grahaṇaviṣayaka	Bsg	19	pañcamyopapada	U5
12	saṅkhyottarapada-vyadhikaraṇa	Bvs	20	saptamyopapada	U7
13	sahapūrvapada-vyadhikaraṇa	BvS	21	mayūravyaṃskādi	Tm
14	prādi-vyadhikaraṇa	Bvp	22	bahupada	Tb
15	upamānapūrvapada-vyadhikaraṇa	BvU	**Karmadhāraya**		
16	nañ	Bsmn	1	viśeṣaṇa-pūrvapada	K1
17	bahupada	Bb	2	viśeṣaṇa-uttarapada	K2
Dvandva			3	viśeṣaṇa-ubhayapada	K3
1	itaretara	Di	4	upamāna-pūrvapada	K4
2	samāhāra	Ds	5	upamāna-uttarapada	K5
3	ekaśeṣa	E	6	avadhāraṇāpūrvapada	K6
anya (others)			7	sambhāvanāpūrvapada	K7
1	dvirukti	d	8	madhyamapadalopi	Km
2	kevala-samāsa	S	-		