An Automatic Sanskrit Compound Processing

Doctor of Philosophy
in
Sanskrit Studies

Anil Kumar
(08HSPH02)

Department of Sanskrit Studies
School of Humanities
University of Hyderabad
Hyderabad

April 2012

An Automatic Sanskrit Compound Processing

A dissertation submitted to the University of Hyderabad
for the award of the degree of

Doctor of Philosophy
in
Sanskrit Studies

by

Anil Kumar
08HSPH02

Under the guidance of
Dr. Amba P. Kulkarni

Department of Sanskrit Studies
School of Humanities
University of Hyderabad
Hyderabad

April 2012

Declaration

I hereby declare that that work embodied in this dissertation entitled "**An Automatic Sanskrit Compound Processing**" is carried out by me under the supervision of *Dr. Amba P. Kulkarni*, Head and Associate Professor, Department of Sanskrit Studies, University of Hyderabad, Hyderabad and has not been submitted for any degree in part or in full to this university or any other university.

Anil Kumar

08HSPH02

Date :

Place : Hyderabad

Department of Sanskrit Studies

University of Hyderabad, Hyderabad

Certificate

This is to certify that Anil Kumar (08HSPH02) has carried out the research-work embodied in the present dissertation entitled "**An Automatic Sanskrit Compound Processing**" at University of Hyderabad. The dissertation represents his independent work and has not been submitted for any research degree of this university or any other university.

Amba P. Kulkarni
Supervisior

Amba P. Kulkarni
Head
Department of Sanskrit Studies

Mohan G. Ramanan
Dean
School of Humanities
University of Hyderabad

Acknowledgements

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study. First and foremost, my utmost gratitude to my supervisor *Dr. Amba P. Kulkarni* for her guidance, patience and support. I consider myself very fortunate for being able to work with a very considerate, dedicated and encouraging person like her. Without her guidance, I would never have been able to finish this work at all. I express my deepest sense of gratitude to her.

I express my sincere gratitude to an iconic and dedictated person *Prof. K. V. Ramakrishnamacharyulu* for his valuable suggestions and guidance throughout my entire research work. I am indebted for his enormous help extended to me.

I express my heartfelt thanks to *Prof. K. N. Murthy* and *Dr. J. S. R. A. Prasad* for their continous support, guidance and encouragements. Their suggestions and encouragements were like vitamins and irons for me which always kept my research work healthy and happy.

My sincere gratitude to *Prof. Oliver Hellwig* and *Mr. Vipul Mittal* who provided me the data and helped me a lot for developing the compound segmenter. I would also like to thank to *Prof. Bhagyalata Pataskar* who provided me various insights regarding compounds.

I express my many thanks to the *Sanskrit Consortium* and *the members of the consortium* who provided me the data for the implementation of Sanskrit compound processor. Without their support, this work was not possible so easily.

I'm indebted to *Dr. Devanand Shukl*, *Mrs. Preeti Shukl*, *Dr. Pankaj Vyas*, *Dr. Sheetal Pokar*, *Dr. Sivaja S. Nair*, *Pavan Kumar* and *Arjun* for their helping during my research. They always provide a support for the in-depth discussions about various research problems, especially *Dr. Devanand Shukl* who gave me various important insights from research point of view.

Without friends the life is incomplete, I thank to all my friends *Siva*, *Surendra*, *Monali*, *Krishna Mohan*, *Gowri* and *Karunakar* for their helping and support.

I am most grateful to my *parents* for their continous support in my every choice. I am also thankful to my wife *Sonia* especially for her support, patience and encouragements. Whenver I need them, they were always available anytime for anything.

I cannot forget office staff of our department, who provided me all kind of intrastructural help. I thank one and all from Department of Sanskrit Studies office.

I would like to thank one and all, who have directly or indirectly been instrumental in the completion of my research work and thesis.

Anil Kumar

Dedicated to my teachers

Prof. K. V. Ramakrishnamacharyulu

and

Dr. Amba P. Kulkarni

Contents

Title Page i

Declaration ii

Certificate iii

Acknowledgements iv

Table of Contents vii

List of Figures xi

1 Overview 1

1.1 Introduction 1

1.2 Goal of research 3

1.3 The organisation of thesis 4

2 Compounds in Sanskrit 8

2.1 वृत्ति 9

2.2 Single wordness (ऐकपद्यम्) 10

2.3 Single accentness (ऐकस्वर्यम्) 10

2.4 Fixed word order of components 11

2.5 Non-insertion of words (अव्यवधान) 12
2.6 Binary formation 12
2.7 Euphonic changes (Sandhi) 13
2.8 Gender 13
2.9 Number 14
2.10 Semantic classification 15
2.11 Syntactic classification 16

3 Semantic Classifications of Sanskrit Compounds 17
3.1 अव्ययीभाव (Endocentric or Indeclinable compound) 18
3.2 तत्पुरुष (Endocentric compound) 20
3.3 बहुव्रीहि 28
3.4 द्वन्द्व (Copulative compound) 31

4 Compound Segmenter 34
4.1 Segmenter 38
4.1.1 Scoring Matrix 38
4.1.1.1 Language Model 39
4.1.1.2 Split Model 40
4.1.2 Segmentation Algorithm 40
4.1.2.1 Results 42

5 Constituency parser 43
5.1 Developing the constituency parser 45
5.2 Statistical approach 45
5.2.1 Base line 46
5.2.2 Our algorithm 47

5.3 Analysis of results 51
5.4 Conclusion 51

6 Sanskrit compound type identifier 53
6.1 Type Identifier 53
6.2 समासविधायकसूत्रs from Pāṇini's Aṣṭādhyāyī 56
6.2.1 अव्ययीभावः 57
6.2.2 तत्पुरुषः 65
6.2.3 बहुव्रीहि 87
6.3 Statistical Apporach 109
6.3.1 Some features of the manually tagged data 110
6.3.2 Algorithm 112
6.3.3 Performance Evaluation 112

7 Sanskrit compound paraphrase generator 115
7.1 Introduction 115
7.2 Paraphrase generator 116
7.2.1 Paraphrase Generation 117
7.2.2 Paraphrase generation of simple compounds 118
7.2.3 Paraphrase generation of nested compounds 121
7.2.4 Problem cases and thier solutions 121
7.3 Evaluation 125

8 Conclusion 126

Appendices 128

A - Table of Compound Paraphrase rules 129

B - Table of Semantic classificatons 130

C - List of aphorisms 132

D - Screenshots of Sanskrit Compound Processor 136

E - Bibliography 141

List of Figures

1.1 Compound Analyser 5

4.1 Compound Splitter : System Data Flow. 41

List of published papers related to thesis

1. Anil Kumar, V. Sheeba and Amba Kulkarni :``*Sanskrit Compound Paraphrase Generator*" in Proceedings of ICON-2009 : 7th International Conference on Natural Language Processing on 14th - 17th December, 09 organised by NLP Association, IIIT, Hyderabad, University of Hyderabad and LDC-IL, CIIL. (Macmillan Publishers, India. ISBN NO. 9780230328457)

2. Anil Kumar, Vipul Mittal and Amba Kulkarni : ``*Sanskrit Compound Processor*" in the proceedings of 4i-SCLS : 4th International Sanskrit Computational Linguistics Symposium on December 10-12, 2010 organised by Special Centre for Sanskrit Studies, Jawaharlal Neharu University, New Delhi. (Springer, 2010, Proceedings : Volume 6465 of Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence Series, ISBN : 3642175279, 9783642175275.

3. Amba Kulkarni and Anil Kumar : ``*Statistical Constituency Parser for Sanskrit Compounds*" in the proceedings of ICON-2011 : 9th International Conference on Natural Language Processing on December 16th -19th, 2011 organised by NLP Association, India, IIIT Hyderabad , LDC-IL, CIIL, Mysore and AU-KBC Centre, Chennai (MACMILLAN : Advanced Research Series, ISBN No. 978-935-059-054-6, published by Macmillan Publishers India LTD)

Chapter 1

Overview

1.1 Introduction

Sanskrit has more than 2500 years old almost exhaustive grammar in the form of Pāṇini's Aṣṭādhyāyī which has the features of computability. However, only recently Sanskrit Computional Linguistics[1] has gained a momentum.

The गणकाष्टाध्यायी[2] software based on Panini's अष्टाध्यायी, provides various search options to get the information with पद-पाठ, अनुवृत्ति, English traslation etc. of rules of अष्टाध्यायी. It also shows the process of generating various nominal declansions and verbal conjugations following Pāṇini's rules.

The French scholar Gérard Huet has done a significant work on the segmentation of Sanskrit texts. His work is purely ruled based. His website

[1] International Symposium on Sanskrit Computational Linguistics held in 2007, 2008, 2009 and 2010

[2] http://www.taralabalu.org/panini/

"Sanskrit Heritage site"[3] provides an automatic Lemmatizer, Tagger, Sanskrit Reader and Sanskrit Parser.

The German scholar Oliver Hellwig has developed a website DCS, the Digital Corpus of Sanskrit[4]. It provides a searching facility for collection of lemmatized Sanskrit texts. It also provides an automatic segmentation and tagging of Sanskrit texts. He follows the statistical methods.

A Special Center for Sanskrit Studies (SCSS) at Jawaharlal Nehru University has developed various tools on Languages processing, Lexical resources, E-learning etc. for Sanskrit under the supervision of Dr. Girish Nath Jha. All the tools are available online at http://sanskrit.jnu.ac.in .

A consortium of Seven institutions[5] has been engaged in developing tools for analysis of Sanskrit. Under this project guidelines and standards for annotation at various levels such as sandhi, samāsa, kāraka, POS etc were developed and a substantial amount of manually tagged data following these standards is generated. The consortium has developed various tools such as sandhi, sandhi-splitter, morphological analyser and generator, POS tagger, sentential parser under this project.

All these tools handle only morphological analysis and segmentation to a large extent. Some of these tools for example Huet's and Hellwig's processors also do sentential parsing. However, there have been almost negligible effort in handling Sanskrit compounds automatically beyond

[3]http://sanskrit.inria.fr/

[4]http://kjc-fs-cluster.kjc.uni-heidelberg.de/dcs/index.php

[5]University of Hyderabad- Hyderabad , Jawaharlal Nehru University- New-Delhi , IIIT-Hyderabad, Sanskrit Academy-Hyderabad, Poornaprajna Vidyaapetha- Bangalore, Rashtriya Sanskrit Vidyapeetha-Tirupati, JRRSU-Jaipur.

segmentation.

A worth noting contribution in the field of Sanskrit compounds is by the Department of Indology of French Institute Pondichery. It has developed a CD version of पाणिनीयोदाहरणकोशः (Volum II - समासप्रकरणम्). It contains a searchable database of compound generation(रुपसिद्धि) of approximately 4,400 compounds from महाभाष्यम्, काशिकावृत्ति, भाषावृत्ति and सिद्धान्तकौमुदी.

On the theoretical side, there are again notable efforts by Gillon and K. V. Ramakrishnamacharyulu towards developing a tagging scheme for compounds. Gillon (Gillon, 2009) suggests tagging of compounds by enriching the context free rules. He does so by specifying the vibhakti, marking the head and also specifying the enriched category of the components. He also points out how certain components such as `na' provide a clue for deciding the type of a compound.

1.2 Goal of research

Sanskrit is very rich in compound formation unlike modern Indian Languages. 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[6]. But mere sandhi splitting does not help a reader in identifying the meaning of a compound, since typically a compound does not code the relation between its components explicitly. To understand the meaning of a compound, it

[6]Sandhi means euphony transformation of words when they are consecutively pronounced. Typically when a word w_1 is followed by a word w_2, some terminal segment of w_1 merges with some initial segment of w_2 to be replaced by a ``smoothed'' phonetic interpolation, corresponding to minimising the energy necessary to reconfigure the vocal organs at the juncture between the words.(Huet, 2006)

is necessary to identify its components and discover the relation between them.

The goal of my research is to evolve a methodology for automatic analysis of Sanskrit compounds. The automatic analysis of a Sanskrit compound involves four major tasks viz. segmentation or component identification, deciding the grouping of components or constituency parsing, identifying the type of a compound and finally paraphrase generation. For building a compound processor, we explore the possibility of both the rule based approach as well as use of statistical methods wherever possible.

1.3 The organisation of thesis

Chapter one (the current chapter) gives a survey of the work in the field of computational linguistics followed by a survey of various works in the field of Sanskrit compounds.

Second chapter gives a brief summary of the features of compounds, the semantics involved, its classification both semantic as well as syntactic. This sets a ground for building the compound processor. We also describe the tagset developed by the Sanskrit Consortium[7] for manual tagging of Sanskrit compounds.

The third chapter describes an architecture of a compound processor. Processing a compound involves four major tasks viz. segmentation (समासपदच्छेदः), constituency parsing (सामर्थ्यनिर्धारणम्), type-identification (समा-

[7]Sanskrit Consortium is a consortium of 7 institutes funded by TDIL programme of DIT for the development of tools for analysis of Sanskrit text and Sanskrit-Hindi Machine Translation.

सभेदनिर्धारणम्) and paraphrase generation (विग्रहवाक्यनिष्पादनम्)(See figure 1.1). These four tasks form the natural modules of a compound processor. The output of one task serves as an input for the next task until the final paraphrase is generated.

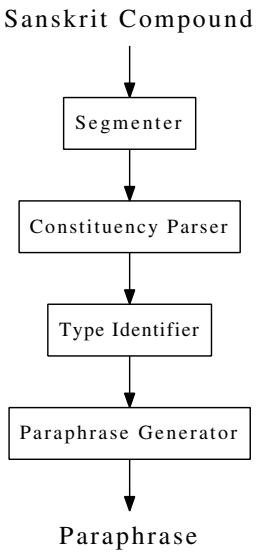


Figure 1.1 : Compound Analyser

The task of a segmenter is to split a compound into its constituents. For instance, the compound

sumitrānandavardhanaḥ

is segmented as

sumitrā-ānanda-vardhanaḥ

Each of the constituent component except the last one is typically a compounding form (a bound morpheme)[8].

[8]with an exception of components of an `aluk' compound.

The constituency parser parses the segmented compound syntactically by pairing up the constituents in a certain order two at a time. For instance,

sumitrā-ānanda-vardhanaḥ

is parsed as

<<sumitrā-ānanda>-vardhanaḥ>

The type-identifier determines the type on the basis of the components involved. For instance,

<<sumitrā-ānanda>-vardhanaḥ>

is tagged as

<<sumitrā-ānanda>T6-vardhanaḥ>T6 .

where T6 stands for compound of type ṣaṣṭī-tatpuruṣa. This module needs an access to the semantic content of its constituents, and possibly even to the wider context.

Finally after the tag has been assigned, the paraphrase generator generates a paraphrase for the compound. For the above example, the paraphrase is generated as :

sumitrāyāḥ ānandaḥ = sumitrānandaḥ,

sumitrānandasya vardhanaḥ = sumitrānandavardhanaḥ.

The fourth chapter describes the first task viz. segmentation. In this chapter, we discuss the computational complexity involved in splitting of a compound into components and then describe how the optimality theory, which is based on Generation-Constraint-Evaluation paradigm, may be used for splitting of compounds. We generate all possible splits following the sandhi rules from Aṣṭādhyāyī and constraint the splits using

morphological analyser. The output is evaluated using simple statistical methods. The algorithm with tested data is explained at the end.

The fifth chapter explains a constituency parser which takes an output of the segmenter and produces a binary tree showing the syntactic composition of a compound corresponding to each of the possible segmentation. In this chapter we describe an algorithm for parsing the components using manually tagged corpus of compounds. We describe how simple conditional probablity may be used for parsing the components. Results of five-fold-testing are given at the end of the chapter.

In the sixth chapter, we deal with an automatic type-identification of compounds. We describe how some of Pāṇini's compound related aphorisms may be used for type-identification. When the rules fail, we use probablity methods to identify the type of compound. We identify the semantic clues for identification of compound type following the aphorisms. We also point out the aphorisms where the clues are difficult to compute or use computationally in the present context.

The seventh chapter is on paraphrase generation, the final task of a compound processor. In this chapter, we give around 55 rules for automatic paraphrase generation corresponding to various sub-types and also describe the algorithm.

The eighth chapter contains the conclusion and provides future directions for improving the automatic sanskrit compound processor further. The use of this analysis for other Indian Languages is also discussed in brief.

Chapter 2

Compounds in Sanskrit

In the world of Natural languages, grammar plays a very important role. Grammar teaches us to use the words in a proper way and it also helps to disambiguate the sentences. In the context of requirement and neccessity of the grammar in languages, पतञ्जलि says in महाभाष्यम् - "रक्षोहाऽऽगमलघ्वसंदेहाः प्रयोजनम्". Word formation is an important part of the grammar. Two distinct phenomenon are involved in word formation. One where new words are formed, and the other one where new base is derived from the old base. The former is known as inflectional morphology and the latter one is called as derivational morphology. Inflectional morpholgy deals with the base forms and affixes and generates the new forms. For instance रामः is generated from the base form राम by adding an inflectional suffix सु. In Derivational morphology, new bases are derived from old base words. For instance दाशरथि (Son of दशरथ) is derived from दशरथ by adding a suffix अण् (in meaning of अपत्यम्). The new base form thus generated denotes a totally new object that is different from the object(s) denoted by the old

base forms. Pāṇini lists 5 different ways of deriving new base forms. They are कृत्-तद्धित-समास-एकशेष-सनाद्यन्तधातु[1]. These are termed as वृत्तिs.

2.1 वृत्ति

The term वृत्ति is derived from the root वृत(वृतु) which means वर्तनम् "the growth" followed by the suffix क्तिन् and it has been defined by पतञ्जलि as परार्थाभिधानं वृत्तिः[2]. It means "The power of expressing a sense different from what was inherent originally in word.". कैयट explains परार्थाभिधानं वृत्ति further as "परस्य शब्दस्य योऽर्थस्तस्याभिधानं शब्दान्तरेण यत्र सा वृत्तिः"[3]. The five वृत्तिs are :

- कृत्- The words or primary nouns are derived from a verbal root. For instance चेयम्, गन्तव्यम्, कारकः etc.
- तद्धित - The secondry nouns derived from Premitive nouns. For instance दाशरथिः "Son of king दशरथ", आश्वपतम् "son of अश्वपति", गार्ग्यः etc.
- समास - The nouns derived by the combination of two or more than two nouns together resulting a single word. For instance घनश्यामः, नीलोत्पलम्, प्रियम्वदा etc.
- एकशेष - Among many similar forms only one form remains and other froms get excluded. एकः शिष्यते अन्यो लुप्यते यत्र is the defintion of एकशेषः from वाचस्पत्यम्. For instance माता च पिता च = पितरौ, सा च सः च = तौ, रामश्च रामश्च = रामौ etc.
- सनाद्यन्तधातु - The derivation of verbal forms by using noun-forms and तिङ् suffixes. For instance पिपठिषति, कृष्णति, बोभूयते etc.

[1]Mentioned by भट्टोजिदीक्षित in सिद्धान्तकौमुदी in समासप्रकरण
[2]In महाभाष्यम् - समर्थः पदविधिः (2.1.1)
[3]In महाभाष्यम्- समर्थः पदविधिः (2-1-1)

Among these वृतिs, समास has a very important and crucial role in word-formation. The compound formation being very productive it forms an open set. Before going deep in compound formation, let us try to understand the meaning of a compound first. The term समास is formed by adding सम् prefix to the अस् root followed by घञ् suffix. Typical Western Linguistics definition of compound is a "lexeme with more than one stems" but in Sanskrit the meaning of compound is described very beautifully by grammarian which is समसनम् इति समासः. that means "the combination of more than one words into one word which conveys the same meaning as that of the collection of the component words together". While combining the components together, a compound undergoes certain operations such as loss of case suffixes, loss of accent, etc..

2.2 Single wordness (ऐकपद्यम्)

A Sanskrit compound is a single word (ऐकपद्यम्[4])(सुप् लोप). Although compound is formed by the combination of more than one word but resultant is a sigle-word i.e. has only one सुप् विभक्ति at the end, with an exception of *aluk* compounds.

2.3 Single accentness (ऐकस्वर्यम्)

The accentuation is the most important part in compound but unfortunately it is lost. Just as compound should be a single word with having single case suffix, compound should also end with a single accent (ऐकस्वर्यम्[5]). Panini has defined three types of accents in grammar : (1) उ-

[4]समर्थः पदविधिः (2.1.1), ऐकपदम् ऐकस्वर्यंञ्च समासवद्भवति - का.-2.1.46

[5]समर्थः पदविधिः (2.1.1), ऐकपदम् ऐकस्वर्यंञ्च समासवद्भवति - का.-2.1.46

दात्त, (2) अनुदात्त and (3) स्वरित. High pitch is known as the उदात्त[6], low pitch is known as the अनुदात्त[7] and high falling pitch is known as स्वरित[8]. While explaining the रक्षोहाऽऽगमलघ्वसंदेहाः प्रयोजनम् [9], Patanjali provides an example of how accent information helps in resolving an ambiguity. For example the word स्थूलपृषती can be identified either as a तत्पुरुष or बहुव्रीहि. However, if we pay attention to the accent, it is not ambiguous. If it has high picth at the end (अन्तोदात्त) then it is तत्पुरुष and if it has pitch in the begining then it is बहुव्रीहि.

2.4 Fixed word order of components

Sanskrit is an almost free word order language. For instance one can write

विद्यायाः पतिः or पतिः विद्यायाः,

कृष्णम् श्रितः or श्रितः कृष्णम्,

पीतम् अम्बरम् यस्य सः or अम्बरं पीतं यस्य सः

However the word order in Sanskrit compounds is fixed. Hence the compounds विद्यापतिः and पतिविद्या have different meanings. So is कृष्णश्रितः and श्रितकृष्णः, or पीताम्बरः and अम्बरपीतः. There are a few exceptions. In कर्मधारय compound, in some cases change of word order is possible.[10] For instance

[6]उच्चैरुदातः(1.2.29)

[7]नीचैरनुदात्तः(1.2.30)

[8]समाहारः स्वरितः(1.2.31)

[9]महाभाष्यम् - पश्पशाह्निके

[10]गुणशब्दयोः समभिव्याहारे विशेषणविशेष्यभावस्य न नियमः । यथा - खञ्जकुब्जः कुब्जखञ्ज इति । क्रियाशब्दयोरप्य-नियमः । यथा - पाचकपाठकः पाठकपाचकः इति । तथा गुणक्रियाशब्दयोरप्यनियमः । यथा - खञ्जपाचकः पाचकखञ्जः इति भाष्ये स्पष्टम् --सि. कौ. - विशेषणं विशेष्येण बहुलम्(2.1.57) - बालमनोरमा टीका ; अत्र हि भाष्यम् - "द्वाविमौ प्रधानशब्दौ (तिलकृष्णौ) एकस्मिन्नर्थे अवरुध्येते । न च द्वयोः प्रधानशब्दयोः एकस्मिन्नर्थे युगपदवरुध्यमानयोः किञ्चिदपि प्रयोजनमस्ति, तत्र प्रयोगादेतत् गन्तव्यं नूनमन्यतरत् प्रधानं तद्विशेषकं चापरमिति । तत्र त्वेतान् सन्देहः किं प्रधानं, कि विशेषणमिति ? स चापि क्व सन्देहः यत्र उभौ गुणशब्दौ ? तद्यथा - खञ्जकुब्जः, कुब्जखञ्ज इति । यत्र हि अन्य तरद् द्रव्यम् अन्यतरो गुणः तत्र

पाचकपाठकः and पाठकपाचकः have the same meanings viz. पाचकश्चासौ पाठकः. Only in the following cases one can change the order of constituents :

1. In case of समानाधिकरण, if both of the constituents are गुणवाची then the order of constituents can be changed. For instance खञ्जकुब्जः[11] can be used as कुब्जखञ्जः.
2. In case of समानाधिकरण, if both of the constituents are क्रियावाची then the order of constituents can be changed. For instance पाचकपाठकः[12] can be used as पाठकपाचकः.
3. In case of समानाधिकरण, if the constituents are like गुण-क्रियावाची then the order of constituents can be changed. For instance खञ्जपाचकः[13] can be used as पाचकखञ्जः.

2.5 Non-insertion of words (अव्यवधान)

Compounds don't allow insertion of word(s) between the components after formation. The compounded form of राज्ञः पुरुषः in राजपुरुषः. One can insert ऋद्धस्य in between राज्ञः and पुरुषः to have राज्ञः ऋद्धस्य पुरुषः. But one can't insert ऋद्धस्य in a compound राजपुरुषः.

2.6 Binary formation

There is a very good discussion in महाभाष्यम् on whether the compound formation is binary or ternary or n-ary. पतञ्जलि argues that the compound formation is binary because the compound formation follows the general rule "सहसुपा" (2.1.4). But there are some cases where more than two

यद् द्रव्यं तत् प्रधानम् । तद्यथा - शुक्लमालभेत कृष्णमालभेत इति ।" -- म०भा० (2.1.57)

[11]सि. कौ. - विशेषणं विशेष्येण बहुलम्(2.1.57) - बालमनोरमा टीका ; म०भा० (2.1.57)

[12]सि. कौ. - विशेषणं विशेष्येण बहुलम्(2.1.57) - बालमनोरमा टीका

[13]सि. कौ. - विशेषणं विशेष्येण बहुलम्(2.1.57) - बालमनोरमा टीका

components are allowed as in द्वन्द्व and बहुपदबहुव्रीहि. It is because, in "अनेक-मन्यपदार्थे" (2.2.24), the word अनेक is used by पाणिनि which is the indication of exceptional of binary for बहुव्रीहि compound and the अनुवृत्ति of the word अनेक goes till "चार्थे द्वन्द्वः" (2.2.29) and that is why we find more than two components in द्वन्द्व compound also.

2.7 Euphonic changes (Sandhi)

संहिता[14] (Proximity) is the basic element which is required for applying Sandhi. Sandhi means "euphony transformation of words when they are consecutively pronounced. Typically when a word w1 is followed by a word w2, some terminal segment of w1 merges with some initial segment of w2 to be replaced by a "smoothed" phonetic interpolation, corresponding to minimizing the energy necessary to reconfigure the vocal organs at the juncture between the words"(Huet G., 2006). The euphonic change (Sandhi) is mandatory in compounds.[15]

2.8 Gender

Constituents of compounds may require a different gender from their original gender due to compound formation. For instance उपगङ्गम् (Near to Ganges river), पाचिकाभार्यः (a man whose wife is a cook) etc.

The compound उपगङ्गम् is an instance of अव्ययीभाव compound where उप is an indeclinable word and the word गङ्गा is a noun with feminine gender. The अव्ययीभाव compounds are typically in neuter gender and hence गङ्गा

[14]"परः सन्निकर्षः संहिता" (1.4.109) ; The closest proximity of letters.

[15]संहितैकपदे नित्या नित्याधातूपसर्गयोः ।
नित्या समासे वाक्ये सा विवक्षामपेक्षते ।।

changes to गङ्गम्. In the same way in पाचिकाभार्यः, the word पाचिका and the word भार्या are feminine in gender. When these are compounded to give a बहुव्रीहि compound returning to a word qualifying a male, it takes masculine gender.[16] In compounds, the last component of compound carries the gender information of the whole compound word as in पाचिकाभार्यः, the word भार्यः is in masculine in gender. It indicates that the gender in compounds can leave their originality according to the context. Panini has listed many aphorisms regarding change of the gender of the last component of compound such as "स नपुंसकम्" (2.4.17), "अव्ययीभावश्च" (2.4.18), "तत्पुरुषोऽनञ्कर्मधारयः" (2.4.19) etc. There is another aphorism given by Panini "परवल्लिङ्गं द्वन्द्वतत्पुरुषयोः" (2.4.26) which denotes that the gender of a द्वन्द्व compound and तत्पुरुष compound is like that of the last word in it. For instance in रामसीते (Rama and Sita), the gender of compound is in feminine in gender. It does not carry the gender of the word राम which is in masculine in gender. All these show that the gender in compounds requires a special treatment, at morphological level.

2.9 Number

The number of a compound depends on the category of compound. If the compound is an अव्ययीभाव then the number of the compound will be singular or if the compound is तत्पुरुष then the number of the last component becomes the number of the whole तत्पुरुष compound. बहुव्रीहि compounds leave their original number and adopt the number of their qualificant. For

16

यल्लिङ्गं यद्वचनं या च विभक्तिः विशेष्यस्य,
तल्लिङ्गं तद्वचनं सा च विभक्तिः विशेषणस्यापि।।

instance पतितपर्णान् वृक्षान् कर्तय (Cut the trees where all the leaves have fallen out.). In this sentence पतितपर्णान् is a बहुव्रीहि compound and it is qualifier of the word वृक्षान्. The word वृक्षान् is in plural. Since the gender, number, case of the qualifier agrees with those of the qualificant, the compound word पतितपर्ण gets the same number and gender of the word वृक्ष.

The number of द्वन्द्व compound depends on the number of components. For instance राधागोविन्दौ, शृङ्गारवीरकरुणाद्भुतहास्यभयानकाः etc. The instance राधागोविन्दौ contains only two components : राधा and गोविन्द so it is in dual number and the instance शृङ्गारवीरकरुणाद्भुतहास्यभयानकाः contains six components : शृङ्गार, वीर, करुणा, अद्भुत, हास्य, and भयानक and hence it gets plural number. There are, however some exceptions. For instance विशश्च शूद्राश्च > विड्शूद्राः[17] (Group of वैश्यs and group of शूद्रs), वृक्षाश्च लताश्च > वृक्षलताः (group of trees and group of branches). In these kind of द्वन्द्व compounds, if the components are related to जातिवाचक or अप्राणिवाचक then the द्वन्द्व compound with two components may be in plural.

2.10 Semantic classification

Semantically पाणिनि classifies the Sanskrit compounds into four major types :

1. अव्ययीभाव : It is an endocentric compound with head typically to the left and behaves as indeclinable. For example उपकृष्णम्, अधिहरि etc.
2. तत्पुरुष : It is en endocentric compound with head typically to the right. For example कृष्णाश्रितः, ग्रामगमी etc.
3. बहुव्रीहि : It is an exocentric compound. For example पीताम्बरः etc.

[17]पा० सू० - जातिरप्राणिनाम् (2-4-6) - सि०कौ०

4. द्वन्द्व : It is a copulative compound. For example रामकृष्णौ, वृक्षलताः etc.

2.11 Syntactic classification

Later grammarian Nagesh Bhatta in his भूषणसार has given the following verse :-

सुपां सुपा तिङा नाम्ना धातुनाऽथ तिङां तिङा ।
सुबन्तेनेति विज्ञेयः समासः षड्विधो बुधैः ।।

The verse indicates six kinds of compounds, classified according to the part of speech of the components involved.

1. सुपां सुपा (Noun with Noun) : This kind of compounds contain only noun constituents. For instance राजपुरुषः.
2. सुपां तिङा (Noun with Verb) : This kind of compounds contain Noun as a first component and conjugated verbs as a second component. For instance पर्यभूषयत्.
3. सुपां नाम्ना (Noun with Stem (Nominal base)) : This kind of compounds contain Noun as a first component and nominal stem as a second component. For instance कुम्भकारः.
4. सुपां धातुना (Noun with Verbal root) : This kind of compounds contain Noun as a first component and Verbal root as a second component. For instance कटप्रूः.
5. तिङां तिङा (Verb with Verb) : This kind of compounds contain only Verbal form as components. For instance पिबतखादता, खादतमोदता.
6. सुपां तिङा (Noun with Verb) : This kind of compound contain Noun as a first component and Verb as a second components. For instance कालविचक्षणा.

Chapter 3

Semantic Classifications of Sanskrit Compounds

Semantically पाणिनि classifies the Sanskrit compounds into four major types :

1. अव्ययीभाव
2. तत्पुरुष
3. बहुव्रीहि and
4. द्वन्द्व

These classifications are not sufficient for generating the paraphrase. For example, the paraphrase of a compound वृक्षमूलम् is वृक्षस्य मूलम् and ग्रामगतः is ग्रामं गतः, though both of them belong to the same class of तत्पुरुष. In the given instances, the paraphrases are different due to the semantic differences and it happens in all the types of compound. Based on their semantic differences, these compounds are further sub-classified into 55 sub-types. All the types and sub-types of compound are described below in this

chapter. The types and sub-types of compound are based on standards evolved by the project entitled "Developmet of Sanskrit Computational Tools and Sanskrit-Hindi Machine Translation System" which is sponsored by Ministry of Information Technology, Government of India, New-Delhi.

3.1 अव्ययीभाव (Endocentric or Indeclinable compound)

अनव्ययम् अव्ययः सम्पद्यते इति is the meaning of अव्ययीभाव which means "In paraphrase, the word which is not the अव्यय but after compound formation that becomes an अव्यय is known as अव्ययीभाव". For instance in उपराजम् (Near to the King), the word राजा is not an अव्यय but when it combines with the word उप which is an अव्यय then the whole word becomes an अव्यय and known as अव्ययीभाव compound. In this type of compound mostly the first member of the compound is predominant[1]. For instance कृष्णस्य समीपम् > उपकृष्णम् (Near to Krishna). In the given instance उप is an indeclinable word and is used in the meaning of सामीप्य (Nearness) and it has primacy in the whole compound.

According to the grammar and by looking the usage of compounds, the अव्ययीभाव compound can be divided into seven sub-types :

1. अव्यय-पूर्वपद-अव्ययीभाव
2. अव्यय-उत्तरपद-अव्ययीभाव
3. तिष्ठद्गुप्रभृति-अव्ययीभाव
4. संख्यापूर्वपद-नद्युत्तरपद-अव्ययीभाव
5. नद्युत्तरपद-अन्यपदार्थसंज्ञायाम्
6. संख्यापूर्वपद-वंश्योत्तरपद-अव्ययीभाव

[1]पूर्वपदार्थप्रधानोऽव्ययीभावः - सि० कौ० - सर्वसमासशेषप्रकरणम्

7. पारे-मध्ये-पूर्वपदषष्ठ्युत्तरपद-अव्ययीभाव

1. अव्यय-पूर्वपद-अव्ययीभाव[2]: The compounds where, the first member of the compound contains an indeclinable word, are known as अव्ययपूर्वपद-अव्ययीभाव compound. For instance उपकृष्णम् (Near to Krishna).
2. अव्यय-उत्तरपद-अव्ययीभाव[3]: In this sub-category, an indeclibable word always occupies second place. For instance सूपप्रति where प्रति is an indeclinable word and stands as a second member of a compound.
3. तिष्ठद्गुप्रभृति-अव्ययीभाव[4] : In this sub-category, the compounds belong the "तिष्ठद्गु" गणपाठ. In तिष्ठद्गु-गणपाठ, the readymade instances of अव्ययीभाव compound are listed by पाणिनि. For instance तिष्ठद्गु, वहद्गु, आयतीगवम् etc.
4. संख्यापूर्वपद-नद्युत्तरपद-अव्ययीभाव[5] : The compounds, where the first component contains a numeral word and the second component contains a word related to rivers, are known as संख्यापूर्वपद-नद्युत्तरपद-अव्ययीभाव compound. For instance सप्तगङ्गम् (the group of seven rivers).
5. नद्युत्तरपद-अन्यपदार्थसंज्ञायाम्[6] : The compounds, where the second component cotains a word related to the rivers and the resulting compound refers to an object other than the referents of the two components in a compound, are known as नद्युत्तरपद-अन्यपदार्थसंज्ञा-अव्ययीभाव compound. For instance उन्मत्तगङ्गम् (the place where the Ganges river becomes very fast).

[2]पा० सू० - अव्ययं विभक्तिसमीपसमृद्धि... (2-1-6)
[3]पा० सू० - सुप् प्रतिनामात्रार्थे (2-1-9)
[4]पा० सू० - तिष्ठद्गु प्रभृतिनि च (2-1-17)
[5]पा० सू० - नदीभिश्च (2-1-20)
[6]पा० सू० - अन्यपदार्थे च सञ्ज्ञायाम् (2-1-21)

6. संख्यापूर्वपद-वंश्योत्तरपद-अव्ययीभाव[7] : The compounds, where the first component stands as a numeral word and second component contains the word related to the वंश्य (one belonging to a family). For instance त्रिमुनि etc.
7. पारे-मध्ये-पूर्वपदषष्ठ्युत्तरपद-अव्ययीभाव[8] : The compounds, where the word पारे (across) and मध्ये (middle) are compounded with a word ending in the first, third and fifth case suffix, are known as पारे-मध्ये-पूर्वपदषष्ठ्युत्तरपद-अव्ययीभाव. For instances पारेगङ्गम् (across the Ganges river), पारेगङ्गात् (across the Ganges river), मध्येगङ्गम् (in the middle of the Ganges river), मध्येगङ्गात् (in the middle of the Ganges river).

3.2 तत्पुरुष (Endocentric compound)

तत्पुरुष compound is an endocentric compound where the second member has primacy[9]. For instance in सूर्यपुत्रः (Son of Sun), the word पुत्रः is the head. The word तत्पुरुष itself is an example of तत्पुरुष compound and it may be paraphrased in several ways as below :

- स चासौ पुरुषः > तत्पुरुषः
- तं पुरुषः > तत्पुरुषः
- तेन पुरुषः > तत्पुरुषः
- तस्मै पुरुषः > तत्पुरुषः
- तस्मात् पुरुषः > तत्पुरुषः
- तस्य पुरुषः > तत्पुरुषः
- तस्मिन् पुरुषः > तत्पुरुषः

[7]पा० सू० - संख्या वंश्येन (2-1-19)
[8]पा० सू० - पारे मध्ये षष्ठ्या वा (2-1-18)
[9]सि० कौ० - उत्तरपदार्थप्रधानस्तत्पुरुषः - सर्वसमासशेषप्रकरणम्

Looking at the विभक्ति (Case marker) involved in these paragraphes one can classify the तत्पुरुष compounds further. In addition there are other kind of तत्पुरुष compounds where the first component is of special type. The grammarians, classify तत्पुरुष compounds further into seven major sub-categories :

1. तत्पुरुष (Determinative compound)
2. कर्मधारय (Descriptive compound)
3. नञ्-प्रादि-कु-गत्यादितत्पुरुष
4. द्विगु
5. उपपदतत्पुरुष
6. मयूरव्यंसकादितत्पुरुष
7. बहुपद-तत्पुरुष

1. तत्पुरुष (Determinative compound) - The first sub-type consists of those तत्पुरुष compound, in which the first word always takes some case-marker in paraphrase. For instances शङ्कुलया खण्डः > शङ्कुलाखण्डः, कृष्णम् आश्रितः > कृष्णाश्रितः etc. This sub-category can then be further divided into seven sub-types on the basis of the case marker the first component takes. These subtypes are : (1) प्रथमातत्पुरुष, (2) द्वितीयातत्पुरुष, (3) तृतीयातत्पुरुष, (4) चतुर्थीतत्पुरुष, (5) पञ्चमीतत्पुरुष, (6) षष्ठीतत्पुरुष and (7) सप्तमी-तत्पुरुष.

   (1)) प्रथमातत्पुरुष[10]- This is an excepitonal compound type among all तत्पुरुष compounds and always contains the first word in nominative case-suffix in paraphrase. For instance उत्तरं कायस्य > उत्तरकायस्य, अर्धं पिप्पल्याः etc.

[10]पाणिनि has given the 6 aphorisms from 2-2-1 to 2-2-6 for this type of compound

(2) द्वितीयातत्पुरुष - The compounds, where a word ending with the second case-suffix is compounded with another nominal word is called द्वितीयातत्पुरुष compound. For instance कृष्णश्रितः etc. Panini has listed out द्वितीयातत्पुरुष compound by few aphorisms such as द्वितीया श्रितातीतपतितगतात्यस्तप्राप्तापन्नैः (2-1-24), स्वयं क्तेन (2-1-25), खट्वा क्षेपे (2-1-26), सामि (2-1-27), कालाः (2-1-28), अत्यन्तसंयोगे च (2-1-28).

(3) तृतीयातत्पुरुष - The compounds, where a word ending with the third case-suffix is compounded with a nominal word is called तृतीया-तत्पुरुष compound. For instance धान्यार्थः etc.

(4) चतुर्थीतत्पुरुष - The compounds, where a word ending with the fourth case-suffix is compounded with a nominal word is called चतुर्थीतत्पुरुष compound. For instance कुण्डलहिरण्यम् etc.

(5) पञ्चमीतत्पुरुष - The compounds, where a word ending with the fifth case-suffix is compounded with a nominal word is called पञ्चमी-त्पुरुष compound. For instance चोरभयम् etc.

(6) षष्ठीतत्पुरुष - The compounds, where a word ending with the sixth case-suffix is compounded with a nominal word is called षष्ठीत-त्पुरुष compound. For instance राजपुरुषः etc.

(7) सप्तमीतत्पुरुष - The compounds, where a word ending with the seventh case-suffix is compounded with a nominal word is called सप्तमीतत्पुरुष compound. For instance अक्षशौण्डः etc.

2. कर्मधारय (Descriptive compound) - The second sub-type consists of those तत्पुरुष compound, in which the components are apposition. The components in this type of compound may be nouns or adjectives qualifying the members. For instances नीलमेघः(blue clouds), पीता-

म्बरः(one who weared yellow cloths) etc. On the basis of semantic differences this category is further divided into 8 sub-divisions : (1) विशेषण-पूर्वपद-कर्मधारय, (2) विशेषण-उत्तरपद-कर्मधारय, (3) विशेषण-उभयपद-कर्मधारय, (4) उपमान-पूर्वपद-कर्मधारय, (5) उपमान-उत्तरपद-कर्मधारय, (6) अवधारणापूर्वपद-कर्मधारय, (7) सम्भावनापूर्वपद-कर्मधारय and (8) मध्यमपदलोपिकर्मधारय.

(1) विशेषण-पूर्वपद-कर्मधारय[11] :- The compounds, where the qualifier stands at the first place and the qualificand stands at the second place, are called as विशेषण-पूर्वपद-कर्मधारय compound. For instance नीलोत्पलम् (the blue lotus).

(2) विशेषण-उत्तरपद-कर्मधारय :- The compounds, where the qualifier stands at the second place and the qualificand stands at the first place, are called as विशेषण-उत्तरपद-कर्मधारय compound. For instance वैयाकरणखसूचिः[12] .

(3) विशेषण-उभयपद-कर्मधारय :- The compounds, where the both the components are adjectives, are called as विशेषण-उभयपद-कर्मधारय compound. For instance मन्दशीतलः, कृताकृतम् [13].

(4) उपमान-पूर्वपद-कर्मधारय[14] :- The compounds, where the first component is found as उपमान (the object of comparison) and the second component is found as सामान्यवचन[15], are called as उपमान-पूर्वपद-कर्मधारय compound. For instance चन्द्रमुखी (the person

[11]पा० सू० - विशेषणं विशेष्येण बहुलम् (2-1-57)

[12]पा० सू० - कुत्सितानि कुत्सनैः(2-1-53)

[13]पा० सू० - क्तेन नञ्विशिष्टेनानञ् (2-1-60)

[14]पा० सू० - उपमानानि सामान्यवचनैः (2-1-55)

[15]उपमानानि सामान्यवचनैरेव समस्यते । उपमानोपमेययोः साधारणः धर्मः सामान्यम् । तद्विशिष्टोपमेयवचनैः समासः इत्यर्थः । मेघश्यामः इत्यत्र श्यामशब्दः सामान्यवचनः । तेन मेघ इव रामः=मेघरामः इति न भवति । - P.7 - समासः, संस्कृतभारती, बेङ्गलूरु

whose face is like Moon).

(5) उपमान-उत्तरपद-कर्मधारय[16] :- The compounds, where the first component is found as उपमेय (the subject of comparison) and the second component is found as उपमान (the object of comparison), are called as उपमान-उत्तरपद-कर्मधारय compounds. For instance पुरुषव्याघ्रः (the person whose is like Tiger), गोवृन्दारकः[17] .

(6) अवधारणापूर्वपद-कर्मधारय :- The compounds, where the उपमान-उपमेय relation is found and identified as a metaphor, are known as अवधारणापूर्वपद-कर्मधारय compound. For instance विद्या एव धनः > विद्याधनः is the treasure.

(7) सम्भावनापूर्वपद-कर्मधारय - The compounds, where a जातिवाचक word is compounded with a वंशवाचक word to which that जातिवाचक belongs, are known as सम्भावनापूर्वपद-कर्मधारय and in this type of compounds, the word related to वंशवाचक is a qualifier and also the first component of the compound. For instance अयोध्यानगरी (the अयोध्या city).

(8) मध्यमपदलोपिकर्मधारय - The compound, where the final member of the first component vanishes while compound formation, is called the मध्यमपदलोपिकर्मधारय or उत्तरपदलोपिकर्मधारय[18] compound. For instance in देवपूजकः ब्राह्मणः, the word पूजक is clearly visible in the paraphrase but in the form of compound the word पूजक vanishes as in देवब्राह्मणः.

3. नञ्-प्रादि-कु-गत्यादितत्पुरुष :-

[16]पा० सू० - उपमितं व्याघ्रादिभिः सामान्यप्रयोगे - (2-1-56)

[17]पा० सू० - वृन्दारकनागकुञ्जरैः पूज्यमानम् (2-1-62)

[18]वा० - शाकपार्थिवादीनां सिद्धये उत्तरपदलोपस्योपसङ्ख्यानम्

- नञ्तत्पुरुष[19] - The compounds where the first component denotes the sense of negation. The नञ्तत्पुरुष compounds are formed by prefixing the particle "न" to another word. In the form of compound it changes to 'अ' before a consonent and to 'अन्' before a vowel. For instance अपर्याप्तम् (not sufficient), अनश्वः (one which is not the horse), अब्राह्मणः (one who is not Brahmin) etc.
- प्रादितत्पुरुष - The compounds where the prepositions including indeclinables are prefixed to another word are called as प्रादित-त्पुरुष[20] compound. For instance प्राचार्यः(eminent teacher), अतिमालः (one who has crossed the river माला) etc.
- कु-तत्पुरुष - The compounds where the first component contains the sense of निन्दा and compounds are formed by prefixing the word 'कु' to anohter word are known as कुतत्पुरुष[21] compound. For instance कुपुरुषः (wicked person), कापुरुषः(wicked person).
- गतितत्पुरुष - The compounds where the preposition are compounded with the primary derivative and indeclinables are called गतितत्पुरुष[22] compound. For instance उरीकृत्वा, उररीकृत्य etc.

4. द्विगु - The fourth sub-type consists of those तत्पुरुष compound, in which the first component contains a numeral adjective and second component contains a noun word, are known as द्विगुतत्पुरुष compound. Three types of द्विगुतत्पुरुष[23] compounds are found in usage : (1) तद्धिता-र्थद्विगु (2) उत्तरपदद्विगु and (3) समाहारद्विगु.

[19]पा० सू० - नञ् (2-2-61)
[20]पा० सू० - कुगतिप्रादयः (2-2-18)
[21]पा० सू० - कुगतिप्रादयः (2-2-18)
[22]पा० सू० - कुगतिप्रादयः (2-2-18)
[23]पा० सू० - तद्धितार्थोत्तरपदसमाहारे च (2-1-51), संख्यापूर्वो द्विगुः (2-1-52)

(1) तद्धितार्थद्विगु - The compounds, which contains the numeral adjectives as first component and a secondry affixed word as second component, are designated तद्धितार्थद्विगु compounds. For instance षाण्मातुरः, द्वैमातुरः etc.

(2) उत्तरपदद्विगु - उत्तरपदद्विगु is the designation of that compound which is compulsorily formed when another word is to be compounded with the so formed द्विगु compound. Thus the compound itself becomes the first member in another compound.[24] For instance पञ्चगवधनः(one whose property consists of five cows.)

(3) समाहारद्विगु - The compounds which suggest the aggregation of a particular thing are known as समाहारद्विगु. For instance पञ्चवटी (the aggregation of five banyan tree), त्रिलोकी (the aggregation of three worlds).

5. उपपद-तत्पुरुष - The fifth sub-type consists of those तत्पुरुष compound, in which the first component contains a उपपद (a noun word) and the second component contains a verbal derivative word (कृत). In this kind of compound the first member can contain any case-relation. For instance कुम्भकारः (The potter) etc. It can further be divided into five more sub-types : (1) द्वितीयोपपद-तत्पुरुष, (2) तृतीयोपपद-तत्पुरुष, (3)चतुर्थ्योपपद-तत्पुरुष (4) पञ्चम्योपपद-तत्पुरुष, (5) सप्तम्योपपद-तत्पुरुष.

(1) द्वितीयोपपद-तत्पुरुष - In this kind of उपपद compound, the first component contains the second-case-marker and the second component contains a verbal derivative word (कृत). For example कुम्भं करोति इति कुम्भकारः, विश्वं पाति इति विश्वपा etc.

[24]Sanskrit compounds-A philosophical study- P.53-Chowkhamba Sankskrit series office, Varanasi

(2) तृतीयोपपद-तत्पुरुष - In this kind of उपपद compound, the first component contains the third-case-marker and the second component contains a verbal derivative word (कृत). For example पद्भ्यां गच्छति इति पद्गः etc.

(3) चतुर्थ्योपपद-तत्पुरुष - In this kind of उपपद compound, the first component contains the fourth-case-marker and the second component contains a verbal derivative word (कृत). For example श्रेयसे तिष्ठति इति श्रेयस्थः etc.

(4) पञ्चम्योपपद-तत्पुरुष - In this kind of उपपद compound, the first component contains the fifth-case-marker and the second component contains a verbal derivative word (कृत). For example दुःखात् जायते इति दुःखजः, शोकात् जायते इति शोकजः etc.

(5) सप्तम्योपपद-तत्पुरुष - In this kind of उपपद compound, the first component contains the seventh-case-marker and the second component contains a verbal derivative word (कृत). For example पङ्के जायते इति पङ्कजः, सरसि जायते इति सरसिजम् etc.

6. मयूरव्यंसकादि-तत्पुरुष[25] - The sixth sub-type is of those तत्पुरुष compound which are known as irregular compounds and found in the मयूरव्यंसकादिगण. For instance मयूरव्यंसकः etc.

7. बहुपद-तत्पुरुष - The seventh sub-type of those तत्पुरुष compound which contains more than two components. For instance द्व्हिसेनम्.

[25]पा० सू० - मयूरव्यंसकादयश्च (2-1-72)

3.3 बहुव्रीहि

बहुव्रीहि compound is an exocentric compound and contains two or more than two components. The first component may contain a noun or an adjective and the second component contains a noun. It becomes an adjective and it adopts the gender and number of the qualificand. For instance रुपवद्भार्यः (one who has a beautiful wife), since the word refers to a man, भार्या a feminine noun changes to भार्य a masculine gendered one. On the basis of semantic differences, it can be divided into two major divisions : (1) समानाधिकरण-बहुव्रीहि and (2) व्यधिकरण-बहुव्रीहि.

1. समानाधिकरण-बहुव्रीहि - This is the designation given to those compounds where all the components contain the same case-endings and same gender. For instance पीतम् अम्बरं यस्य सः > पीताम्बरः, आरूढः वानरः येन सः > आरूढवानरः, रूपवती भार्या यस्य सः > रूपवद्भार्यः etc. It is of 11 types : (1) द्वितीयार्थबहुव्रीहि, (2) तृतीयार्थबहुव्रीहि, (3) चतुर्थ्यर्थबहुव्रीहि, (4) पञ्चम्यर्थबहुव्रीहि, (5) षष्ठ्यर्थबहुव्रीहि, (6) सप्तम्यर्थबहुव्रीहि. (7) दिग्वाचक-बहुव्रीहि, (8) संख्योभयपद-बहुव्रीहि, (9) उपमानपूर्वपद-बहुव्रीहि, (10) प्रहरणविषयक-बहुव्रीहि and (11) ग्रहणविषयक-बहुव्रीहि :-

   (1) द्वितीयार्थबहुव्रीहि :- The बहुव्रीहि compounds, which contain the second case suffix, come under this type of compound. For instance प्राप्तम् उदकम् यं सः > प्राप्तोदकः .

   (2) तृतीयार्थबहुव्रीहि :- The बहुव्रीहि compounds, which contain the third case suffix, come under this type of compound. For instance ऊढः रथः येन सः > ऊढरथः.

   (3) चतुर्थ्यर्थबहुव्रीहि :- The बहुव्रीहि compounds, which contain the fourth case suffix, come under this type of compound. For instance दत्तं

वस्त्रं यस्यै सा > दत्तवस्त्रा.

(4) पञ्चम्यर्थबहुव्रीहि :- The बहुव्रीहि compounds, which contain the fifth case suffix, come under this type of compound. For instance पतितानि पर्णानि यस्मात् सः > पतितपर्णः.

(5) षष्ठ्यर्थबहुव्रीहि :- The बहुव्रीहि compounds, which contain the sixth case suffix, come under this type of compound. For instance पाचिका भार्या यस्य सः > पाचिकाभार्यः.

(6) सप्तम्यर्थबहुव्रीहि :- The बहुव्रीहि compounds, which contain the seventh case suffix, come under this type of compound. For instance वीराः पुरुषाः यस्मिन् सः > वीरपुरुषः.

(7) दिग्वाचक-बहुव्रीहि :- The बहुव्रीहि compounds, which contain names of directions, are designated as दिग्वाचक-बहुव्रीहि compound. For instance पूर्वस्याः उत्तरस्याः च दिशो यदन्तरालम् > पूर्वोत्तरा.

(8) संख्योभयपद-बहुव्रीहि :- The बहुव्रीहि compounds, where both the components are numeral words, are known as संख्योभयपद-बहुव्रीहि compound. For instance द्वौ वा त्रयः वा > द्वित्राः.

(9) उपमानपूर्वपद-बहुव्रीहि :- The बहुव्रीहि compounds, where the first component conatains a उपमानवाचक word, are called as उपमानपूर्वपद-बहुव्रीहि compound. For instance चन्द्र इव मुखं यस्याः सा > चन्द्रमुखी.

(10) प्रहरणविषयक-बहुव्रीहि :- The बहुव्रीहि compounds, where both the components alike in form (and represented by the word-pattern) तत्र (or) तेन (are compounded with each other, the compound) conveying the meaning इदम् "this". For instance द-ण्डादण्डी.

(11) ग्रहणविषयक-बहुव्रीहि :- The बहुव्रीहि compounds, where both the

components alike in form (and represented by the word-pattern) तत्र (or) तेन (are compounded with each other, the compound) conveying the meaning इदम् "this". For instance केशाकेशि.

2. व्यधिकरण-बहुव्रीहि - व्यधिकरणबहुव्रीहि is the designation given to those types of compounds where all the components do not contain the same case and gender. For instance ईश्वरे निष्ठा यस्य सः > ईश्वरनिष्ठः. In this instance, the word ईश्वर is in seventh case-suffix and the word निष्ठा contains first case-suffix. There is no rule given by पाणिनि to identify the conflicted case suffix in व्यधिकरणबहुव्रीहि compound. For instances

गदा पाणौ यस्य सः > गदापाणिः

कण्ठे काल यस्य सः > कण्ठेकालः

भाले चन्द्रः यस्य सः > भालचन्द्रः

विषं कण्ठे यस्य सः > विषकण्ठः

It can also be divided into six types of व्यधिकरण-बहुव्रीहि compound : (1) सङ्ख्योत्तरपद-व्यधिकरण-बहुव्रीहि, (2) सहपूर्वपद-व्यधिकरण-बहुव्रीहि, (3) प्रादि-व्यधिकरण-बहुव्रीहि, (4) उपमानपूर्वपद-व्यधिकरण-बहुव्रीहि, (5) नञ्-बहुव्रीहि and (6) बहुपद-बहुव्रीहि.

(1) सङ्ख्योत्तरपद-व्यधिकरण-बहुव्रीहि :- The बहुव्रीहि compounds where the second member of compound contains a numeral word and first member may contain an indeclinable word, are designated as सङ्ख्योत्तरपद-व्यधिकरण-बहुव्रीहि. For instance दशानां समीपे ये सन्ति ते > उपदशाः.

(2) सहपूर्वपद-व्यधिकरण-बहुव्रीहि :- The बहुव्रीहि compounds where second component is prefixed by the word 'स' or 'सह' are known as सहपूर्वपद-व्यधिकरण-बहुव्रीहि compound. For instance सपुत्रः or सहपुत्रः.

(3) प्रादि-व्यधिकरण-बहुव्रीहि :- The बहुव्रीहि compounds where the first component

contains prefixe including indeclinables are know as प्रादि-बहुव्रीहि compounds. For instance निर्गता दया यस्मात् सः > निर्दयः.

(4) उपमानपूर्वपद-व्यधिकरण-बहुव्रीहि :- The बहुव्रीहि compounds, where the first component conatains a उपमानवाचक word, are called as उपमानपूर्वपद-व्यधिकरण-बहुव्रीहि compound. For instance उष्ट्रस्य इव मुखं यस्य सः > उष्ट्रमुखः.

(5) नञ्-बहुव्रीहि - The बहुव्रीहि compounds where the first component contains the prefix the particle "न". For instance अविद्यमानः पुत्रः यस्य सः > अपुत्रः, अविद्यमानम् अपत्यं यस्य सः > अनपत्यः etc.

(6) बहुपद-बहुव्रीहि - The बहुव्रीहि compounds where more than two components are found, are called as बहुपदबहुव्रीहि compounds. For instance पञ्चगवधनः, पूर्वशालप्रियः etc.

3.4 द्वन्द्व (Copulative compound)

द्वन्द्व compound is a copulative compound and contains two or more than two components. द्वन्द्व compound is also an exception of binary. In this compound, the components may have nouns or adjectives. The gender of the last component becomes the gender of the whole द्वन्द्व compound. The attribute 'च' connects all components into a paraphrase. The main purpose of construction of द्वन्द्व compound is to show the aggregation of individuals. For instance रामसीतालक्ष्मणभरतशत्रुघ्नाः अयोध्यां गच्छन्ति (राम, सीता, लक्ष्मण, भरत and शत्रुघ्न are going to Ayodhya city.), In this sentence रामसीतालक्ष्मणभरतशत्रुघ्नाः is a द्वन्द्व compound formed from the five padas राम, सीता, लक्ष्मण, भरत and शत्रुघ्न and all of them are going to Ayodhya city.

On the basis of Semantic differences, द्वन्द्व compound can be divided into three types : (1) इतरेतर-द्वन्द्व, (2) समाहार-द्वन्द्व and (3) एकशेष-द्वन्द्व.

(1) इतरेतर-द्वन्द्व :- The द्वन्द्व compound, where all the components are independent and predominent, are designated as इतरेतर-द्वन्द्व compound. The number of a द्वन्द्व compound depends on the number of components and the gender of a द्वन्द्व compound depends on the gender of last component of the compound. For instance रामसीते (Rama and Sita), व्याकरणन्यायमीमांसाशास्त्राणि (व्याकरण-शास्त्रम् and न्याय-शास्त्रम् and मीमांसा-शास्त्रम्) etc.

(2) समाहार-द्वन्द्व :- When द्वन्द्व compound gives the sense of aggregation of individuals then it becomes the समाहार-द्वन्द्व compound and the aggregation of individals is the main purpose here, so it gets neuter gender for whole compound and the number of this compound becomes singular. For instance संज्ञा च परिभाषा च एतयोः समाहारः > संज्ञाप-रिभाषम् (The aggregation of संज्ञा and परिभाषा).

(3) एकशेष-द्वन्द्व :- एकशेष-द्वन्द्व is an exceptional compound. एकः शिष्यते अन्यो लु-प्यते यत्र is the characterisation of this compound. For instances पितरौ. The word पितरौ means माता च पिता च (Mother and Father) but in the compound formation only the last component remains.

Apart from these types and categories two more types of compounds also have been found in usages :- (1) केवलसमास and (2) द्विरुक्तिसमास.

1. केवल-समास[26] is an excepitonal compound and found rarely in usages. For instance पूर्वं भूतो भूतपूर्वः.
2. द्विरुक्ति-समास is a kind of compound where the first component occures twice and frequently found in usages. For instance उपर्युपरि, अधोऽधः etc.

[26]तत्पुरुषादिसंज्ञाविनिर्मुक्त : समाससंज्ञामात्रयुक्त : केवलसमास : । अर्थात् यस्य समासस्य नास्ति नाम कश्चित् स : समास : केवलसमास : इति अभिधीयते (ज्ञायते) ।

This classification as mentioned above is based on the standards developed by the Sanskrit Consortium[27]. This classifications needs further refinement. For exampls उपपदतत्पुरुष compound and व्यधिकरण-बहुव्रीहि compound may be further classified into more sub-types depending on the विभक्ति the first component takes.

[27]"Developmet of Sanskrit Computational Tools and Sanskrit-Hindi Machine Translation System" which is sponsored by Ministry of Information Technology, Government of India, New-Delhi.

Chapter 4

Compound Segmenter[1]

The task of a compound segmenter is to split a compound into its constituents. The segmentation of a compound requires the sandhi rules and a morphological analyser which can validate the splits. If we see Panini's Astadhyayi, we find the rules only for sandhi synthesizing not for splitting and sandhi is mandatory in compounds. The compound segmenter uses the same sandhi rules for splitting. Sandhi rules are in the form of triples (x,y,z) where x is the last letter of the first component and y is the first letter of second component. z contains the euphonic changes of x+y. For instance in the compound word सूर्योदयः (The rising of Sun), the word सूर्य is the first component and the word उदयः is the second component. The letter 'अ' is the last letter of the word सूर्य which is x and the letter 'उ' is the first letter of the word उदयः which is y. The z is x+y which is 'ओ'. To split a sandhi thus we need the sandhi rules in reverse form, viz given z, what are the possible values of x and y. The sandhi

[1]This work is jointly done with Vipul Mittal, an MS student of IIIT-Hyderabad.

rules are deterministic, i.e, given x and y, z is unique [with an exception of some विभाषा, where sandhi is optional, and thus x and y are unchanged in these cases]. These sandhi rules when inverted, lead to non-determinism. To illustrate, consider the following sandhi rules

a + a -> ā

a + ā -> ā

ā + a -> ā

ā + ā -> ā

When inverted, given 'ā', now there are 4 different ways of splitting it. All these possible splits should be constrained further to be morphologically valid. As an example following these rules, the string 'tatrāpi' can be split in 4 different ways as

tatra + api

tatra + āpi

tatrā + api

and tatrā + āpi

Of these only the first one will be validated by the morphological analyser and the other three answers wil be discarded. The above two words viz tatra and api are padas themselves and in case of tatrāpi mere sandhi has taken place. This is not an example of compound.

Issues in identifying the components of a compound :

When a compound is split, among the resulting components only the last one has a vibhakti and hence can be recognised by a morphological

analyser. But what about the components in the initial position? The compound रामालयः, suppose split as राम-आलयः, आलयः will be recognised by the morphological analyser. राम will also be analysed as a संबोधन form of राम. But as is obivious, in case of रामालयः, राम is not in the संबोधन form. Here राम is in its compound initial form (समासपूर्वपद). There are various similar issues related to identification of components of a compound, which are described below.

(a) **Identification of** समासपूर्वपद :- The component undergoes various morphological changes when it is used as a समासपूर्वपद. These changes are :

(i) Deletion of न् from प्रातिपदिक :- The 'न्' at the end of a प्रातिपदिक gets deleted when used as a पूर्वपद. For instance राजन् becomes राज in राजपुत्रः.

(ii) ह्रस्वविधान - In some cases the पूर्वपद changes to ह्रस्व. For instances the word इष्टका becomes इष्टक in इष्टकचितम् [2].

(iii) दीर्घविधान - In some cases the पूर्वपद gets दीर्घ. For instance the word भ्रु becomes भ्रू in भ्रूकुटिः[3]

(iv) आदेशविधान - In some cases the पूर्वपद gets substitute by another word. For instace the word अङ्गुलि becomes अङ्गुल in अङ्गुलाकर्णः, the word हृदय becomes हृद् in हृल्लेखः etc.

(b) **Identification of** समास-उत्तरपद :-

(i) **Bound morphemes** :- In case of उपपद-तत्पुरुष, we come across bound morphemes such as कारः, ज्ञः, जः etc as in कुम्भकारः, स-

[2] "इष्टकेषीकामालानां चिततूलभारिषु" (6-3-65)

[3] "इको ह्रस्वोऽङ्योः गालवस्य" (6-3-61) - सि० कौ०

र्वज्ञः, अग्रजः etc. The components कारः, ज्ञः, जः are not पदs, they do not have an independent exisence. But in case of a compound identification, these are to be recognised. कार, ज्ञ, ज being the प्रातिपदिकs, take all the possible vibhaktis and thus follow the regular paradigms.

(ii) **Change in gender** :- Two types of compounds viz the अव्ययीभाव and बहुव्रीहि may advocate gender changes in the final components. For example in उपगङ्गम्, the noun गङ्गा which is in feminine gets neuter gender due to the अव्ययीभाव compound formation. Similarly in case of पीताम्बरः, the second compound अम्बरः is in masculine gender whereas the gender of अम्बर is neuter.

(iii) **Change in number**:- Consider the word अनेकान्. This will be split by a compound splitter as अन्-एकान्. Here the पूर्वपद अन् should be recognised as a variation of न and एकान् should be recognised morphologically. Though एक is used many senses, in this context एक is used to indicate the संख्या(number) and hence can't have एकान् form which is in plural. But it needs to be recognised. So this is to be treated as a special case.

(iv) **Change in paradigms** :- In case of न अस्ति किञ्चन यस्य = अकिञ्चनः[4], we require a new paradigm to recognise the word किञ्चन.

To handle these phenomenon, a seperate module for morphological analysis is added. This module is invoked only on समासपूर्वपद and समासोत्तरपदs avoding the overgeneralisation.

[4]पा०सू० - ``मयूरव्यंसकादयश्च'' (2-1-72)

4.1 Segmenter

The task of a compound segmenter is to split a compound into its constituents. The segmentation of a compound requires the sandhi rules and a morphological analyser which can validate the splits. The segmenter uses reversed sandhi rules and first produces all the possible splits. These splits are then validated by morphological analyser and then the correct splits are selected. This filter produces multiple possibilities. These need to be ranked so that most probable answers are easily accessible. We describe below the GENerate-CONstrain-EVALuate cycle of the segmenter attributed to the optimality theory.

4.1.1 Scoring Matrix

A parallel corpus of Sanskrit text in sandhied and unsandhied form is being developed as a part of the Sanskrit Consortium project in India. The corpus contains texts from various fields ranging from children stories, dramas, purāṇas to Ayurveda texts. From around 100K words of such a parallel corpus, 25K words were found to be in sandhied forms. These 25K parallel instances of sandhied and unsandhied text were extracted and were used to get the frequency of occurence of various sandhi rules. If no instance of a sandhi rule is found in the corpus, for smoothing, we assign the frequency of 1 to this sandhi rule.

We define the estimated probability of the occurrence of a sandhi rule as follows :

Let R_i denote the ith rule with f_{R_i} as the frequency of occurrence in the

manually split parallel text. The probability of rule R_i is :

$$P_{R_i} = \frac{f_{R_i}}{\sum_{j=1}^{n} f_{R_j}}$$

where n denotes the total number of sandhi rules found in the corpus.

Let a word be split into a candidate S_j with k constituents as $< c_1, c_2, ..., c_k >$ by applying $k-1$ sandhi rules $< R_1, R_2, ..., R_{k-1} >$ in between the constituents. It should be noted here that the rules $R_1, ..., R_{k-1}$ and the constituents $c_1, ..., c_k$ are interdependent since a different rule sequence will result in a different constituents sequence. The sequence of constituents are constrained by a language model whereas the rules provide a model for splitting. We define two measures each corresponding to the constituents and the rules to assign weights to the possible splits.

4.1.1.1 Language Model

Let the unigram probability of the sequence $< c_1, c_2, ..., c_k >$ be PL_{S_j} defined as :

$$PL_{S_j} = \prod_{x=1}^{k} (P_{c_x})$$

where P_{c_x} is the probability of occurrence of a word c_x in the corpus.

4.1.1.2 Split Model

Let the splitting model PS_{S_j} for the sandhi rules sequence $< R_1, R_2, ..., R_{k-1} >$ be defined as :

$$PS_{S_j} = \prod_{x=1}^{k-1} (P_{R_x})$$

where P_{R_x} is the probability of occurrence of a rule R_x in the corpus.

Therefore, the weight of the split S_j is defined as the product of the language and the split model as :

$$W_{S_j} = \frac{PL_{S_j} * PS_{S_j}}{k}$$

where the factor of k is introduced to give more preference to the split with less number of segments than the one with more segments.

4.1.2 Segmentation Algorithm

The approach followed is GENerate-CONstrain-EVALuate. In this approach, all the possible splits of a given string are first generated and the splits that are not validated by the morphological analyser are subsequently pruned out. Currently we apply only two constraints viz.

- C1 : All the constituents of a split must be valid morphs.
- C2 : All the segments except the last one should be valid compounding forms.

The system flow is presented in Figure 4.1.

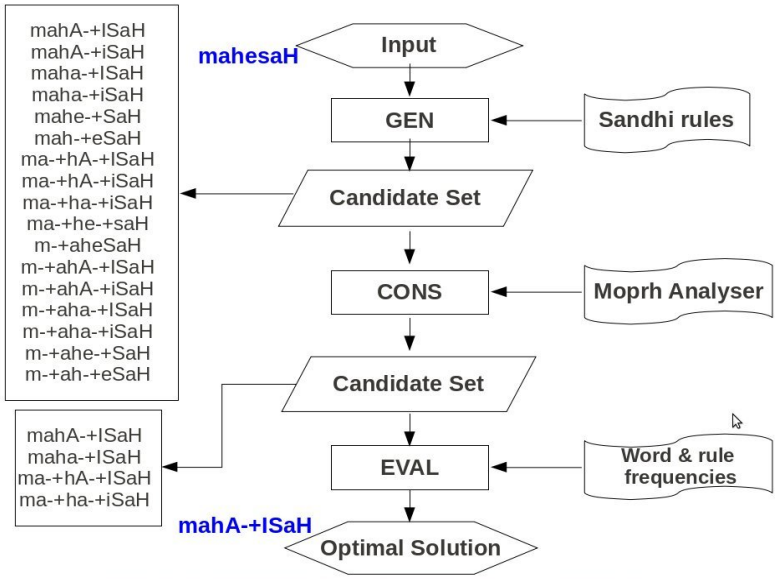


Figure 4.1 : Compound Splitter : System Data Flow.

The basic outline of the algorithm is :

1. Recursively break a word at every possible position applying a sandhi rule and generate all possible candidates for the input.
2. Pass the constituents of all the candidates through the morph analyser.
3. Declare the candidate as a valid candidate, if all its constituents are recognised by the morphological analyser, and all except the last segment are compounding forms.
4. Assign weights to the accepted candidates and sort them based on the weights as defined in the previous subsection.
5. The optimal solution will be the one with the highest weight.

4.1.2.1 Results

The current morphological analyser[5] can recognise around 140 million words. Using 2,650 rules and a test data of around 8,260[6] words parallel corpus for testing, we obtained the following results :

- Almost 92.5% of the times, the first segmentation is correct. And in almost 99.1% of the cases, the correct split was among the top 3 possible splits.
- The precision was about 92.46% (measured in terms of the number of words for which first answer is correct w.r.t. the total words for which correct segmentation was obtained).
- The system consumes around 0.04 seconds per string of 15 letters on an average.[7]

The complete rank wise distribution is given in Table 4.1.2.1.

Rank	% of words
1	92.4635
2	5.0492
3	1.6235
4	0.2979
5	0.1936
>5	0.3723

Rank-wise Distribution

[5]available at http://sanskrit.uohyd.ernet.in/scl/morph/index.html.
[6]The test data is extracted from manually split data of Mahābhāratam.
[7]Tested on a system with 2.93GHz Core 2 Duo processor and 2GB RAM.

Chapter 5

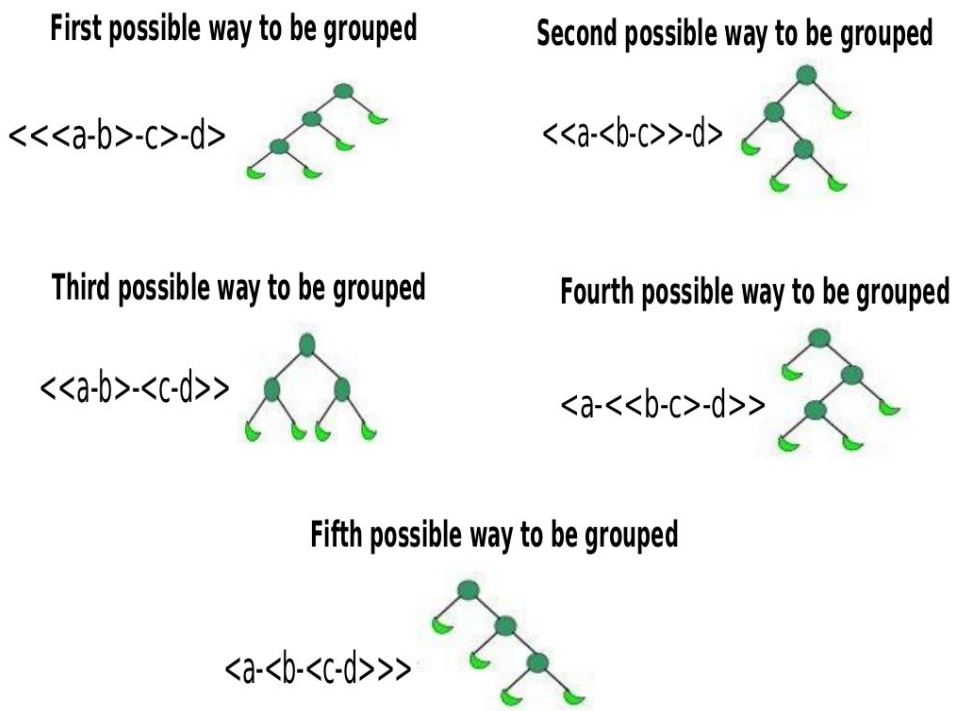
Constituency parser

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 show 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, 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 as illustrated by the following two examples.

1. < eka - < priya - darśanaḥ >>
   (Gloss : <one - <who is dear to all>>)
2. << tapas-svādhyāya > - niratam >
   (Gloss : <<Penance-self-study>-constantly engaged>)

With 3 components, only these two parses are possible. But as the number of constituents increase, the number of possible ways the constituents can

be grouped grows very fast. For instance a compound with 4 components *a-b-c-d* can be grouped into 5 possible ways. See the figure below.



The 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) where for $n \geq 0$,

$$C_n = \frac{(2n)!}{(n+1)!n!}$$

5.1 Developing the constituency parser

We observed that the correctness of parse is governed by the semantics. It is the semantic compatibility (सामर्थ्य) between two words which decides the parse of a compound. We found that there are two approaches to decide semantic compatibility (समार्थ्यनिर्धारण) :-

1. Use the semantically rich lexicon and follow a ruled based approach.
2. Use statistical patterns of manually parsed compounds to decide the more likely parse.

We follow the second approach to parse the compounds.

5.2 Statistical approach

The task of the constituency parser is then to choose the correct way of grouping the components together, in a given context. The meaning compatibility among the components rule out the possibility of most of the parses, eventually leading to one or may be a small number of possible parses. The Sanskrit Consortium has developed a manually tagged Sanskrit corpus of around 600K words, which has around 80K instances of compound words. These compounds are split into components and also tagged and parsed manually. Table 1 describes the corpus statistically.

The compounds with more than 2 components need parsing. Though Sanskrit is a free word order language, the components in a compound have a fixed word order, and they also show natural tendency towards left branching. In other words, in case of a compound with 3 components, the number of compounds with <a-< *b*-*c* >> pattern were less compared to <<a-*b* >-c>. The manually tagged data supports with this observation.

Table 5.1 : Statistical details of corpus

Total size in words	150K
Total compounds	30K
Compounds with 2 components	21,384
Compounds with 3 components	6,809
Compounds with 4 components	1,321
Compounds with 5 components	319
Compounds with more than 5 components	133

Table 2 and 3 show the number of occurences of different parsed structures with 3 components and 4 components.

Table 5.2 : Compounds with 3 components

Pattern no.	Patterns	No. of instances	Instances (in %)
1	<a-<b-c>>	466	6.8
2	<<a-b>-c>	6343	93.2

Table 5.3 : Compounds with 4 components

Pattern no.	Patterns	No. of instances	Instances (in %)
1	<a-<b-<c-d>>>	5	0.3
2	<<a-<b-c>>-d>	127	9.7
3	<a-<<b-c>-d>>	33	2.3
4	<<<a-b>-c>-d>	832	63
5	<<a-b>-<c-d>>	324	24.7

5.2.1 Base line

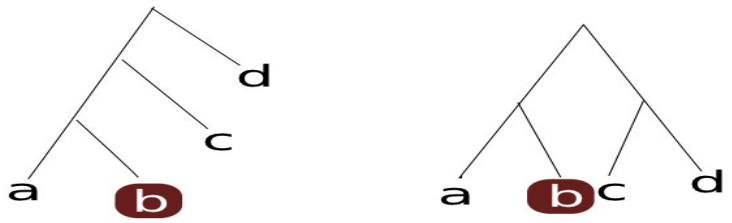
As is clear from Table 2, the data is skewed and even a simple algorithm assigning the most frequent pattern will result into 93% and 63% accuracy in case of compounds with 3 components and 4 components respectively.

5.2.2 Our algorithm

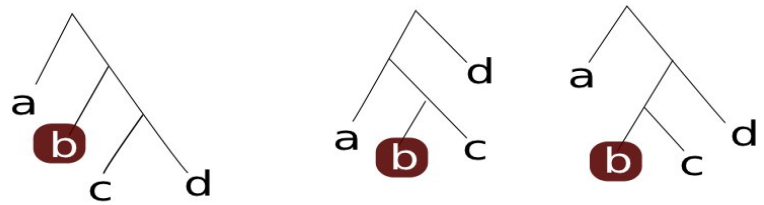
The main task here is to decide the compatibility (sāmarthya or yogyatā) between the two words. The conditional probabilities are used to decide the compatibility between a pair of words. Let us take an example of 3 component compound viz a-b-c. Now, to decide the parse of this compound essentially means to decide whether the component `b' joins with `a' or with `c'. In a compound <a-b>, the component `a' is termed as iic (in initio compositi or samāsa pūrvapada) and the component `b' is termed as ifc (in fine compositi or samāsa uttarapada). Thus to decide the parse of a compound a-b-c, one should know whether it is more likely to use `b' as an iic or ifc refer the figure below.

However, only the unigram frequencies to this effect are not sufficient, since the context, or the affinity of other words viz. `a' and `c' in the context plays a role in determining the parse. For examples in case of four components as a-b-c-d, the component b is a final component in two parses and initial component in three parses as shown in figure -

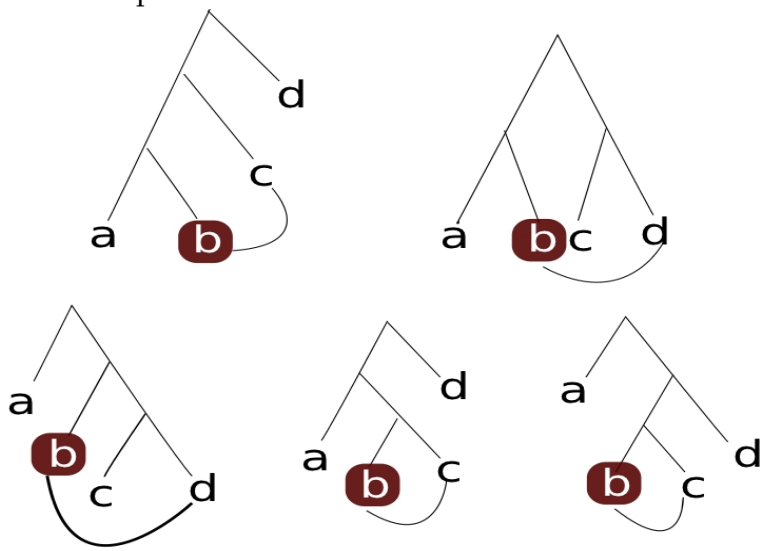
b as final component :



b as initial component :



So to fix the parse, we need one more parameter viz whether 'b' is related to 'c' or 'd', in other words whether 'b' has more affinity towards 'c' or 'd', and this fixes the parse as shown below -



So what is needed is the comparison between two conditional probabilities viz. the probability of `b' as an ifc given that `a' is an iic and the probability of `b' as an iic given that `c' is the ifc. If the difference between these conditional probabilities is above a certain threshold, we use this information to decide the parse. When the conditional probabilities are not available or if the difference between the two conditional probabilities is not significant, we resort to the unigram frequencies, and based on the probabilities of `b' as ifc versus iic, we take the decision.

Thus the algorithm may be summarised as :

let p(ab) = probability of b as ifc given a is iic,
let p(bc) = probability of b as iic given c is ifc,
let p(bf) = probability of b as ifc,
let p(bi) = probability of b as iic,
and let T = threshold.
if((p(ab) - p(bc)) > T) then the parse = <<a-b>-c>
else if((p(bc) - p(ab)) > T) then the parse = <a-<b-c>>
else if (p(bf) - p(bi) > T) then the parse = <<a-b>-c>
else if (p(bi) - p(bf) > T) then the parse = <a-<b-c>>
else the default parse = <<a-b>-c>

We used the manually tagged corpus for training as well as testing. The data was randomised and split into 5 sets, and a 5-fold testing method was adopted, eachtime reserving one set for testing and using other 4 sets for

training. The average results of this experiment using the above algorithm are shown in Table 4 below.

Table 5.4 : Performance of compounds with 3 components

Patterns	No. of Instances	Pecision (in %)	Recall (in %)	F-measure
<a-<b-c>>	471	57	45	0.503
<<a-b>-c>	6408	95.8	97.27	0.965

The average performance is 93.66% which is very close to the baseline, and at the same time the F-measure for the frequent pattern is also close to 1.

This algorithm was extended to more than 3 components, and the performance for 4 components is shown in the Table 5.

Table 5.5 : Performance of compounds with 4 components

Patterns	No. of Instances	Pecision (in %)	Recall (in %)	F-measure
<a-<b-<c-d>>>	5	50	20	0.28
<<a-<b-c>>-d>	127	80	3	0.57
<a-<<b-c>-d>>	33	-	0	-
<<<a-b>-c>-d>	832	72	87	0.788
<<a-b>-<c-d>>	324	79	40	0.53

The average performance for 4 components is 65.4% which is again just above the base line performance.

Examples with more than 4 component compounds being very small in number, their precision and recall are not measured. The overall performance for 5 test data is given below.

The average success rate is 86.5%.

Sr no.	Total compounds	Correct instances	Wrong instances	% success
1	1738	1493	245	85.9
2	1734	1503	231	86.6
3	1729	1497	232	86.5
4	1759	1532	227	87.0
5	1737	1500	237	86.3

5.3 Analysis of results

Analysis of wrong results of 3 component compounds is carried out manually to understand the reasons behind failure. The failures were of two types.

(a) Where the instances of <<a-b>-c> type were parsed by machine as <a-<b-c>>, in most of these cases `a' was an adjective of `b'. Since the evidences of <b-c> were found in the corpus, whereas the evidences of <a-b> were not found, machine did not have the information that `a' is an adjective of `b'. Hence these compounds were wrongly classified as `<a-<b-c>>'.

(b) The cases where the instances of <a-<b-c>> were classified as <<a-b>-c>, as such instances of <b-c> were not found in the training data and machine produced the default left branching parse viz <<a-b>-c>.

5.4 Conclusion

1. The $iifs$ or the final components in the compound are inflected. Instead of the inflected words, if the probabilities are calculated

taking into account the roots, the performance should increase.

2. The compounds are of 4 different types with head being different in each of them. The algorithm described in section 4, should take into account the compound type to decide its association. For example, consider the compound <<a-b >-c >where <a-b >is an avyayībhāva (left word to be the head), then it is not 'b' which will be associated with 'c' but it is 'a' which will be associated with 'c'. The current implementation since does not take into account the type of a compound, it produces wrong grouping in such cases. The algorithm needs to be thus modified to take into account the compound type as well along with the association.
3. The components of compound together convey a unique meaning which is over and above its components meanings. So had the components of a Sanskrit be written seperately, they are very close to a Multi Word Expression (MWE). So the problem of constituency parsing of Sanskrit compounds then is directly relevent for determining the association of words within a MWE with each other. The insights gained in handling compounds will be directly available for handling MWEs of Indian languages.

Chapter 6

Sanskrit compound type identifier

6.1 Type Identifier

After getting a constituency parse of a compound, the next task in the compound analysis is to assign an appropriate operator (tag) to each non-leaf node. This operator then operates on the leaf nodes to produce the associated meaning. For instance,

<<sumitrā-ānanda>-vardhanaḥ>

is tagged as

<<sumitrā-ānanda>T6-vardhanaḥ>T6

Tag specifies the relation between the components. This relation is semantic in nature and is not expressed by any morpheme. So it is only the semantics of the components involved that determines this relation. Hence unless one knows the पदार्थs (the meaning of a word) of the components, one can't determine the meaning of a compound. That means deciding the relation between the components is not easy. There are two basic questions one needs to answer.

1. What is the nature of the relations ?
2. Where are the clues that help in deciding the relations ?

Below we give two examples that help answer these questions. Consider the compounds राजपुरुषः, दशरथपुत्रः and वृक्षशाखा. In the first case the relation between राज and पुरुषः is that of servant-master (सेव्यसेवकभावः), in the second it is father-son (पिता-पुत्रसम्बन्धः) and in the third case it is part of (अवयव-अवयविसम्बन्धः). However, in all the three cases instead of specifying these deeper relations, the relation between the components is expressed by the genetive case suffix in the paraphrase of these compounds as राज्ञः पुरुषः, दशरथस्य पुत्रः and वृक्षस्य शाखा, and hence these are classified as षष्ठीत-त्पुरुष. So, for deciding the relation between components, instead of deeper semantic analysis, we resort to the Paninian classification of compounds. At the same time it was observed earlier the coarse grain distinction of compounds into four major classes is not sufficient. We need the fine grain distinction, as specified in Appendix - B.

Now consider another compound रामेश्वरः, depending on the context this may mean

a) रामश्चासौ ईश्वरश्च

b) राम ईश्वरः यस्य सः

c) रामस्य ईश्वरः

In the first case it is कर्मधारय, in the second case it is बहुव्रीहि and in the third it is षष्ठी तत्पुरुष. Though the semantic content of the components is the same in all the three cases, the relations between these components is determined by other words in the context.

Thus the compound type is guided by the paraphrase of a compound, the paraphrase depends on the context, and the present day computer technology is still not fully equipped to handle context. So on the face of it the problem seems to be difficult to handle. However a close look at the Pāṇini's sūtras dealing with compounds, provide us a lot of semantic clues. For example, look at the following sūtras

"द्वितीया श्रितातीतपतितगतात्यस्तप्राप्तापन्नैः" (2.1.24)

"तृतीया तत्कृतार्थेन गुणवचनेन" (2.1.30)

"अन्नेन व्यञ्जनम्" (2.1.34)

"पञ्चमी भयेन" (2.1.37)

Each of these gives a criterion either in terms of semantics or as a list of possible components in a compound of a particular type. This helps in deciding the compound types. However, such rules do not cover all the cases. Only for a few type of compounds, which is typically a closed list, the rules are provided. To decide the type in all other cases, we need to look at the semantics of the components involved.

There are two possible ways to handle the semantics. The first one is to use semantically tagged lexicon and build a rule-base for identification. Another approach is to use manually tagged corpus and machine learning algorithms to develop automatic identifier. In what follows we first look at all the relevent sūtra from Pāṇini and discuss the feasibility of their implementation. Next we discuss the building of automatic tagger using the manually tagged corpus and some simple statistical insights.

6.2 समासविधायकसूत्रs from Pāṇini's Aṣṭādhyāyī

In the whole Aṣṭādhyāyī the process of compound formation is optional and comes under the sūtra विभाषा "2-1-11" That is the speaker has an option to use either a paraphrase such as राज्ञः पुरुषः or a compounded form राजपुरुषः to express the same meaning. There are a few exceptions to it viz. यथाशक्ति.

Given a paraphrase Pāṇini describes the process of compound formation. To be precise, the process starts with the अलौकिक-विग्रह and not the लौकिक-विग्रह. Then the order of the components is decided followed by its type, the process then deals with the elision of the vibhakti of intermediate components, assiging the svaras, addition of the certain suffixes to change the gender if neccessary and finally the compound form is generated.

The सूत्रs dealing with compounds then can be broadly classified into two types

a) सूत्रs that deal with the semantic content to decide the type of a compound.

b) सूत्रs that deal with the process of compound formation involving

   (i) deciding the word order

   (ii) deletion of विभक्ति

   (iii) assigning a स्वर

The second type of सूत्रs are thus useful from generation point of view. They deal with the morpholgy and phonology. The first type of सूत्रs provide semantic clues for deciding the type of a compound. See the Appendix-C. We describe below whether each of these सूत्रs is useful in deciding the type of a compound, followed by the difficulties in using some of

them computationally. To explain the aphorisms below, we have reffered the books "The Astadhyayi of Panini"[1] and "The Siddhanta Kaumudi of Bhattoji Dixita"[2].

6.2.1 अव्ययीभावः

1. सूत्र - "अव्ययं विभक्ति-समीप-समृद्धि-व्यृद्ध्यर्थाभावात्यया-सम्प्रति-शब्दप्रादुर्भाव..." - 2.1.6

   Meaning - An indeclinable (is invariably compounded with a case-inflected word) conveying the meaning of (1) a case suffix, (2) nearness, (3) prosperity, (4) misfortune, (5) absense of a thing, (6) passing (of a particular time), (7) not (the proper time) now, (8) reputation of a name, (9) after, (10) yathā, (11) one coming after the other, (12) simultaneity, (13) resemblance, (14) fittingness, (15) totality, (16) end (the compound being called avyayībhāva).

   Description - This is a vidhisūtra as well as a sajñāsūtra. This aphorism provides a list of semantic clues to the initial component of a compound of type अव्यय-पूर्वपद-अव्ययीभाव. Below we give a list of examples.

   (1) विभक्ति (a case suffix) - अधिहरि is an example of विभक्तिवचन and हरौ इति is its paraphrase. Here अधि is an indeclinable word and it forces the word हरि to be into seventh case-suffix.

   (2) सामीप्य (nearness) - उपराजम् is an instance of सामीप्य. Here the word उप indicates the meaning 'nearness' to राजा and it also indicates that the second component will contain the sixth case-suffix. राज्ञः

[1]Vol-V and VI, written by S. D. Joshi and JAF. Roodbergen, Published by Sahitya Academy

[2]Srisa Chandra Vasu, Published by Motilal Banarssidas publishers

समीपम् is the paraphrase of उपराजम्.

(3) समृद्धि (prosperity) - मद्राणां समृद्धिः is the paraphrase of सुमद्रम् and it is derived in the sense of समृद्धि. Here सु is an indeclinable word.

(4) व्यृद्धि (misfortune) - दुर्यवनम् and दुर्गवादिकम् are the instances of व्यृद्धि. Here दुर् is an indeclinable word and it is used in the sense of misfortune.

(5) अभाव (absense of a thing) - निर् is an indeclinable which is mostly used in the sense of absense. For instance निर्मक्षिकम् means मक्षिकाणाम् अभावः (absense of flies).

(6) अत्यय (passing (of a particular time)) - the word अत्यय means passing, departure or destruction of a particular time. For instance निर्हिमम् means हिमस्य अत्ययः (the snows have passed).

(7) असम्प्रति (not (the proper time) now) - The word असम्प्रति suggests a time which is not proper to do. For instance अतितैसृकम्, derived in the sense of उपभोगस्य वर्तमानकालप्रतिषेधः 'the rejection of the present for use'. The काशिकावृत्ति explains that तैसृक means अच्छादन 'cloak'. So the compound means that now is not the proper time to wear a cloak.

(8) शब्दप्रादुर्भाव (reputation of a name) - इतिपाणिनि derived in the sense of प्रकाशता शब्दस्य (the fame of a name). Here the word पाणिनि is a reputed name.

(9) पश्चात् (after) - In the sense of after indeclinables such as अनु are used. For instance अनुविष्णु means विष्णोः पश्चात् "after the Vishnu".

(10) यथार्थ (yathā) - The word यथा has various meanings such as योग्यता(Capability), वीप्सा(pervasion), पदार्थानतिवृत्ति(not going beyond

the word's meaning) and सादृश्यम्(Similarity). For instances अनु-रूपम्, प्रत्यर्थम्, यथाशक्ति etc.

(11) आनुपूर्व्य (one coming after the other) - As अनुज्येष्ठं प्रविशन्तु भवन्तः (let your honors enter in the order of seniority)

(12) यौगपद्य (simultaneity) - The word सचक्रम् is derived in the sense of चक्रेण युगपत् or सहचक्रेण.

(13) सादृश्य (resemblance) - ससखि, derived in the sense of सदृशः किख्यः 'one similar to a monkey'. Why mention सादृश्य as a separate category, if सादृश्य has already been stated as one of the meaning of यथा? The idea is that by this separate statement of सादृश्य compound-formation can also be justified when the notion of सादृश्य is not the main one, but the subordinate one.

(14) सम्पत्ति (fittingness) - सब्रह्म 'in a way fitting for a Brahmin'.

(15) साकल्य (totality) - सतृणम् अभ्यवहरति 'He eats right down to the grass'.

(16) अन्त (end) - साग्नि अधीते 'he studies up to the end of (the section on) अग्नि'.

Clue for the type identification - Though it is not possible to decide the type of compounds in all the cases listed above, still a few can be classified automatically. If the initial component is an indeclinable such as प्र, परा, अप, सम्, अनु etc. and the final component is in neuter gender then the compound may be an अव्यय-पूर्वपद-अव्ययीभाव.

2. सूत्र - "यावदवधारणे" (2-1-8)

Meaning - (the indeclinable) यावत् in the sense of अवधारण

(specification) (is invariably compounded with a case-inflected word the compound being called अव्ययीभाव.)

Description - The current aphorism requires the context of अवधारण (limitations) to decide the type of an अव्ययीभाव compound. For instance यावच्छ्लोकम् (यावन्तः श्लोकाः तावन्तः अच्युतप्रणामाः).

Clue for the type identification - If the पूर्वपद is यावत् and उत्तरपद is in neuter gender then the compound may be an अव्यय-पूर्वपद-अव्ययीभाव compound, and such compounds typically carry the meaning अवधारण.

3. सूत्र - "सुप्प्रतिना मात्रार्थे" (2-1-9)

Meaning - A case inflected word (is invariably compounded with the indeclinable) prati in the sense of mātrā "a small quantity" (being called अव्ययीभाव).

Description - A word ending in a case-affix is compounded with the word प्रति in the context of mātrā (a small quantity) and the compound is called an अव्ययीभाव.

Clue for the type identification - If the उत्तरपद is प्रति the compound may be an अव्यय-उत्तरपद-अव्ययीभाव and the meaning of प्रति is a small quantity. These are very rare and the only instances we know of are सूपप्रति "a small quantity of soup" and शाकप्रति "a small quantity of vegetables".

4. सूत्र - "अक्षशलाकासंख्याः परिणा" (2-1-10)

Meaning - (The case inflected words) akṣa "(cubic) die", śalākā

"(oblong) die" and numerals (are invariably compounded) with pari. Like previous aphorism (2-1-9), this is also a special rule of अव्ययीभाव. The compound is confined to words denoting loss in gambling (being called अव्ययीभाव).

Examples - अक्षपरि (an unlucky throw of dice), शलाकापरि (an unlucky throw of oblong), एकपरि (an unlucky throw by one over).

Clue for the type identification - If the पूर्वपद contains the words such as अक्ष, शलाका or numerals like एक,द्वि etc. and the उत्तरपद contains an indeclinable word परि then the compound may be of type अव्यय-उत्तरपद-अव्ययीभाव.

5. सूत्र - "अपपरिबहिरञ्चवः पञ्चम्या" (2-1-12)

Meaning - (The case-inflected words) apa, pari, bahis and (those ending in) -अञ्च् are marginally (compounded with (a word ending in) an ablative case ending (being called अव्ययीभाव).

Examples - अपविष्णु संसारः or अप विष्णोः संसारः (The world is outside or away from Viṣhṇu). So also बहिर्ग्रामम् or बहिर्ग्रामात् (outside the village) etc.

Clue for the type identification - If पूर्वपद contains अप, परि and बहि or any word ending in -añc and the उत्तरपद contains either fifth case-affix or is of nueter gender then the compound is called अव्यय-पूर्वपद-अव्ययीभाव.

Problem - Here, it is not easy to know whether the initial word ends in अञ्च् or not unless we have the derivational information of all the head words in the lexicon. However, if such a rich lexicon is

available, then automatic identification is possible.

6. सूत्र - "आङ् मर्यादाऽभिविध्योः" (2-1-13)

Meaning - (The case-inflected word) आङ् in the sense of मर्यादा (limit exclusive) or अभिविधि (limit inclusive) (is marginally compounded with a case-inflected word (being called अव्ययीभाव).

Examples - आपाटलिपुत्रं or आपाटलिपुत्रात् वृष्टो देवः (It rained up to Pātaliputra), आकुमारम् or आकुमारेभ्यः यशः पाणिनेः (The fame of Pāṇini extends even to the boys) etc.

Clue for the type identification - If the पूर्वपद contains the indeclinable आङ् and the उत्तरपद is either in neuter gender or fifth case-affix then the compound is an अव्यय-पूर्वपद-अव्ययीभाव.

7. सूत्र - "लक्षणेनाऽभिप्रती आभिमुख्ये" (2-1-14)

Meaning - (the case-inflected words) अभि and प्रति in the sense of आभिमुख्य "direction towards" (are marginally compounded) with (a case-inflected word indicating) the goal (the compound when formed, conveying the same meaing as the formally corresponding wordgroup and being called अव्ययीभाव).

Example - अभ्यग्नि शलभाः पतन्ति "the moths fall in the direction of fire" etc.

Clue for the type identification - If the first component is either अभि or प्रति, this is अव्यय-पूर्वपद-अव्ययीभाव. However, it is not possible to decide the meaning of अभि or प्रति.

8. सूत्र - "अनुर्यत्समया" (2-1-15)

   Meaning - (the case-inflected word) अनु (is marginally compounded with a case-inflected word standing for an indicator) nearness to which (is conveyed, the compound when formed, conveying the same meaning as the formally corresponding wordgroup and being called अव्ययीभाव).

   Example - अनुवनमशनिर्गतः "the thunder bolt fell near the forest" etc.

   Clue for the type identification - As above, if the first component of the compound is अनु then it may be classified as अव्यय-पूर्वपद-अव्ययीभाव. However, it is not possible to determine the meaning of अनु mechanically.

9. सूत्र - "यस्य चायामः" (2-1-16)

   Meaning - also (the case-inflected word अनु is marginally compounded with a case-inflected word standing for an indicator) whose length (is conveyed, the compound, when formed, being called अव्ययीभाव).

   Examples - अनुगङ्गं वाराणसी "the city of Banaras extends alongside the river Ganges", अनुयमुनं मथुरा "Mathura alongside of the Yamuna, the length of Mathura being measured by that of the Yamuna".

   Clue for the type identification - The clue, as in the above rule, is in the first component अनु and the type of the compound is अव्यय-पूर्वपद-अव्ययीभाव compound. However, meaning of अनु can not be determined mechanically.

10. सूत्र - "तिष्ठद्गुप्रभृतीनि च" (2-1-17)

    Meaning - This is a विधिसूत्र which assigns the designation अव्ययीभाव to specifically enumerated words (being called अव्ययीभाव).

    Description - By this aphorism Panini mentions all irregular forms of अव्ययीभाव compounds. All the words are listed in गणपाठ. For instance तिष्ठद्गु (at milking time) etc. These are all treated as special cases and handled separately.

11. सूत्र - "पारे मध्ये षष्ठ्या वा" (2-1-18)

    Meaning - (the case-inflected words) pāre (across) and madhye (in) are preferably compounded with (a case-inflected word) ending in a genitive case ending (being called अव्ययीभाव).

    Examples - पारेगङ्गात् or गङ्गापारात् (bring across the Ganges), मध्येगङ्गात् or गङ्गामध्यात् (middle of the Ganges) etc.

    Clue for the type identification - If पूर्वपद contains the words पारे or मध्ये and the उत्तरपद contains the fifth case-affix then the compound is called पारे-मध्ये-पूर्वपद-षष्ठ्युत्तरपद-अव्ययीभाव. Similarly if second component contains either पारात् or मध्यात् then also the compound is called पारे-मध्ये-पूर्वपद-षष्ठ्युत्तरपद-अव्ययीभाव.

12. सूत्र - "संख्या वंश्येन" (2-1-19)

    Meaning - A numeral is preferably compounded with a case-inflected word signifying one who belongs to a family (being called अव्ययीभाव).

    Examples - त्रिमुनि, द्विमुनि etc.

Clue for the type identification - In this kind of compound, the first component contains a numeral and the second component is a वंश्यवाची. To implement this aphorism one has to provide the lexicon with वंश्यवाची tag. Such a compound will be tagged as संख्यापूर्वपद-वंश्योत्तरपद-अव्ययीभाव.

13. सूत्र - "नदीभिश्च" (2-1-20)

Meaning - Also (a numeral is preferably compounded with case-inflected words) signifying a river (being called अव्ययीभाव).

Examples - द्वियमुनम्, सप्तगङ्गम् etc.

Clue for the type identification - To implement this aphorism, one has to enrich the lexicon with the names of the rivers. If the first component of compound contains a numeral and the second component contains the नदीवाचक words, then it is of type संख्यापूर्वपद-नद्युत्तरपद-अव्ययीभाव compound.

6.2.2 तत्पुरुषः

14. सूत्र - "द्वितीया श्रितातीतपतितगतात्यस्तप्राप्तापन्नैः" (2-1-24)

Meaning - (a word ending in) the second case-termination (is preferably compounded) with (the case-inflected words) श्रित "who has fallen upon", अतीत "who has passed beyond", पतित "who has fallen into", गत "who has gone to", अत्यस्त "who has thrown beyond", प्राप्त "who has attained" and आपन्न "who has obtained" (being called तत्पुरुष).

Description - This is a विधिसूत्र prescribing the preferred formation of

an accusative तत्पुरुष compound with क्तान्त forms conveying an active meaning. The examples quoted by काशिकावृत्ति are :

1. कष्टश्रितः 'who has landed in trouble', derived in the sense of कष्टं श्रितः.
2. कान्तारातीतः 'who has passed beyond the forest', derived in the sense of कान्तारम् अतीतः.
3. नरकपतितः 'who has fallen into hell', derived in the sense of नरकं पतितः.
4. ग्रामगतः 'who has gone to his village', derived in the sense of ग्रामं गतः.
5. तरङ्गात्यस्तः 'who has thrown beyond the waves', derived in the sense of तरङ्गान् अत्यस्तः.
6. सुखप्राप्तः 'who has attained happiness', derived in the sense of सुखं प्राप्तः.
7. सुखापन्नः 'who has obtained happiness', derived in the sense of सुखम् आपन्नः.

Clue for the type identification - If the stem of the second component of a compound is from the list mentioned above viz. श्रित, अतीत, पतित etc. then this is an instance of द्वितीया-तत्पुरुष compound.

15. सूत्र - "स्वयं क्तेन" (2-1-25)

Meaning - (the case-inflected word) स्वयं "(by) oneself" (is preferably compounded) with a case-inflected word) ending in क्त (being called तत्पुरुष).

Description - This is a विधिसूत्र prescribing तत्पुरुष compound-formation. The indeclinable word स्वयम् 'oneself', is compounded with a word ending in the affix क्त, and the resulting compound is

called तत्पुरुष. The anuvritti of the word द्वितीया being inappropriate does not take place; though, however it is needed in the next aphorism. Because the word स्वयम् being an indeclinable, cannot take any case-affix. As स्वयं कृतस्यापत्यं = स्वायङ्कृतिः.

Clue for the type identification - If the first component contains an indeclinable स्वयं or its variation स्वायम् and the second component contains an inflected क्तान्त word then it is an instance of तत्पुरुष compound.

16. सूत्र - "खट्वा क्षेपे" (2-1-26)

Meaning - (the case-inflected word) खट्वा "bed" (ending in the second case-termination is compounded with a case-inflected word ending in the suffix क्त) when (the sense of) abuse is to be conveyed (the compound being called तत्पुरुष).

Example - खट्वारूढः जाल्मः (literally lying on a bed); silly, stupid, going wrong or astray.

Clue for the type identification - If the first component contains the word खट्वा and the second component is an inflected क्तान्त then we may postulate that it is likely to be a द्वितीयातत्पुरुष compound. The censure context is neccessary for confirmation, and mechanical detection of censure context being impossible, we can just mark this as a possibility and can not confirm the type.

17. सूत्र - "सामि" (2-1-27)

Meaning - (the case-inflected word) सामि "incompletely, half" (is

preferably compounded with a case-inflected word ending in क्त (being called तत्पुरुष).

Description - This is a विधिसूत्र prescribing a तत्पुरुष compound-formation. The indeclinable word सामि meaning 'half' is compounded with a word ending in the affix क्त, and the resulting compound would be तत्पुरुष. As सामिकृतम् 'half-done'.

Clue for the type identification - If the first component contains an indeclinable सामि and the second component contains a क्तान्त word then it is a तत्पुरुष compound.

18. सूत्र - "कालाः" (2-1-28)

Meaning - (a case-inflected word ending in the second case-termination) signifying (a period of) time (is preferably compounded with a case-inflected word ending in क्त (being called तत्पुरुष).

Description - This is विधिसूत्र prescribing तत्पुरुष compound formation. The words denoting the time, (but not the duration thereof) being in the accusative case, are optionally compounded with a word ending in the suffix क्त, and the resulting compound is तत्पुरुष.

Example - मासप्रमितश्चन्द्रमाः 'the new moon' (literally, the moon that has begun to measure the month).

Clue for the type identification - For handling this type of compound, the list of the words which denote the time is required. If such a word is the first component and the second component contains the क्तान्त word and it may be classified as तत्पुरुष.

19. सूत्र - "पूर्वसदृशसमोनार्थकलहनिपुणमिश्रश्लक्ष्णैः" (2-1-31)

Meaning - (a case-inflected word ending in the third case-termination is preferably compound) with (case-inflected words) conveying the sense of पूर्व "previous", सदृश "like", सम "same", ऊन "less", and with (the case-inflected words) कलह "fight", निपुण "skilful", मिश्र "mixed" and श्लक्ष्ण "smooth" (being called तत्पुरुष).

Examples - From this aphorism we learn incidentally that the words पूर्व etc. govern the instrumental case. The examples quoted by the काशिकावृत्ति are :-

(a) मासपूर्वः derived in the sense of मासेन पूर्वः, who is earlier by (one) month.

(b) मातृसदृशः 'resembling his mother', पितृसदृशः 'resembling his father'.

(c) मातृसमः 'equal to his mother'.

(d) माषोनम् 'what is short by (one) माष.

(e) असिकलहः 'a fight with swords', वाक्कलहः 'a fight with words'.

(f) वान्निपुणः 'skilful with words', आचारनिपुणः 'skilful in behaviour'.

(g) गुडमिश्रः 'mixed with molasses'.

(h) आचारश्लक्ष्णः 'gentle in behaviour'.

Clue for the type identification - The current aphorism gives the list of the possible second components and in such cases it may be तृतीयातत्पुरुष compound.

20. सूत्र - "चतुर्थी तदर्थार्थबलिहितसुखरक्षितैः" (2-1-36)

Meaning - (a case-inflected word) ending in the fourth case-termination (is preferably compounded) with (a case-inflected word

signifying) a thing for the sake of that (meaning expressed by the word in the fourth case) and with (the case-inflected words अर्थ "thing", बलि "food-offering", हित "benificial", सुख "pleasant" and रक्षित "reserved" (being called तत्पुरुष).

Description - By the expression तदर्थ 'for the purpose there of', the special relation of a material and its modification alone is, held to be intended.[3] This we infer from the ज्ञापक of the words बलि and रक्षित used in the aphorism. As यूपाय दारुः = यूपदारुः 'wood for stake' (that is wood which by modification will be changed into a stake). But not so here. रन्धनाय स्थाली 'pot for cooking'. So also or अवहननायोलुखलम् 'the wood mortar for threshing'.

Form this aphorism we may also infer that the Dative case conveys also the sense of 'for the purpose there of'. The word तदर्थार्थ is taken as one word by some, meaning 'a thing serviceable there to'.

Examples - भूतबलिः 'a sacrifice for भूतs', गोहितम् 'what is good for cows', गोसुखम् 'what is pleasant for cows' and गोरक्षितम् 'what is kept for cows'.

Clue for the type identification - The current aphorism gives the list of words for second components. If we find any one of the word from the list as a second component then we may assume that the given compound is a चतुर्थीतत्पुरुष compound.

21. सूत्र - "पञ्चमी भयेन" (2-1-37)

Meaning - (a case-inflected word) ending in the fifth case ending (is preferably compounded) with (the case-inflected word) भय "fear"

[3]चतुर्थी तदर्थमात्रेण चेत्सर्वप्रसङ्गः। सर्वस्य चतुर्थ्यन्तस्य तदर्थमात्रेण सह समासः प्राप्नोति। अनेनापि प्राप्नोति - रन्धनाय स्थाली, अवहननायोलुखलमिति। किं कारणम् ? अविशेषात् । न हि कश्चिद्विशेष उपादीयते - एवंजातीयकस्य चतुर्थ्यन्तस्य तदर्थेन सह समासो भवति इति। अनुपादीयमाने विशेषे सर्वप्रसङ्ग।

(being called तत्पुरुष).

Examples - चोरात् भयम् = चोरभयम् 'fear from thieves'.

Clue for the type identification - The current aphorism gives the information regarding the second component. In the compound, if the second component contains the words भय, भीति or भी then the compound type can be identified as a पञ्चमीतत्पुरुष.

22. सूत्र - "अपेतापोढमुक्तपतितापत्रस्तैरल्पशः" (2-1-38)

Meaning - 'in a few instances a case-inflected word ending in the fifth case-termination is preferably compounded) with (the case-inflected words) अपेत "parted from", अपोढ "caried off", मुक्त "loosened from", पतित "fallen from" and अपत्र "recoiling from" (being called तत्पुरुष).

Examples - सुखापेतः 'gone away from pleasure', कल्पनापोढः 'carried away beyond imagination', चक्रमुक्तः 'freed from the wheel', स्वर्गपतितः 'fallen from heaven', तरङ्गापत्रस्तः 'afraid of the waves'.

Why is the word अल्पशः 'when the action is gradual' used ? It shows the limited range of this kind of compounds. Not every ablative word can be so compounded. Hence, there is no compounding at all in the following case :- प्रसादात् पतितः, 'fallen from the mansion'. For here, the fall, is violent and sudden and not gradual and slight.

Clue for the type identification - The current aphorism gives the information regarding the second component of a तत्पुरुष compound. In a compound, if we get one of these words then the compound type can be identified as a पञ्चमीतत्पुरुष.

23. सूत्र - "स्तोकान्तिकदूरार्थकृच्छ्राणि क्तेन" (2-1-39)

Meaning - (case-inflected words ending in the fifth case-termination) conveying the sense of स्तोक 'a little', अन्तिक 'near' or दूर 'far' and (the case-inflected word ending in the fifth case-termination) कृच्छ्र 'difficulty' (are preferably compounded) with (a case-inflected word derived with the suffix) क्त (being called तत्पुरुष).

Examples - स्तोकान् मुक्त 'loosed from a little distance', अन्तिकादागतः 'come form near', अभ्याशादागतः 'come from near', दूरादागतः 'come form far', विप्रकृष्टादागतः 'come from a distance', कृच्छ्रादागतः 'come with difficulty'.

Clue for the type identification - The current aphorism explains a अलुक्समास and gives a list of first components in a तत्पुरुष compound. In a compound, if we get one of these words on fifth case of as a first component and the second component is a क्त ending कृदन्त then the compound type can be identified as a पञ्चमीतत्पुरुष.

24. सूत्र - "सप्तमी शौण्डैः" (2-1-40)

Meaning - (a case-inflected word) ending in the seventh case-termination (is preferably compounded) with (the case-inflected words) शौण्ड 'cunning', etc (being called तत्पुरुष).

Example - अक्षशौण्डः 'skilled in dice'.

The current aphorism explains a सप्तमीतत्पुरुष compound and gives a list of second component. In the aphorism the word शौण्डैः suggests a शौण्डादि-गण(group) from गणपाठ and the words are :- 1) शौण्ड 2) धूर्त 3) कित 4) व्याड 5) प्रवीण 6) संवीत 7) अन्तर 8) अधि 9) पटु 10) पण्डित 11) कुशल 12)

चपल 13) निपुण 14) संव्याड 15) भन्य and 16) समीर.

Clue for the type identification - In a compound, if the second component is from शौण्डादिगण then the compound type may be identified as a सप्तमीतत्पुरुष.

25. सूत्र - "सिद्धशुष्कपक्वबन्धैश्च" (2-1-41)

Meaning - 'also (a case-inflected word ending in the seventh case-termination is preferably compounded) with (the case-inflected words) सिद्ध 'well-known', शुष्क 'dried', पक्व 'cooked' and बन्ध 'binding'(being called तत्पुरुष).

Examples - साङ्काश्यसिद्धः 'perfect in साङ्काश्य', आतपशुष्कः 'dried in sun', स्थालीपक्वः 'cooked in pot' and चक्रबन्धः 'bound on the wheel'.

Clue for the type identification - The current aphorism gives a list of second components. In a compound, if the second component contains one of these words then the compound type may be identified as सप्तमीतत्पुरुष.

26. सूत्र - "ध्वाङ्क्षेण क्षेपे" (2-1-42)

Meaning - (a case-inflected word ending in the seventh case-termination is compounded) with (the case inflected word) ध्वाङ्क्ष 'crow' when (the compound conveys) the sense of abuse (the compound being called तत्पुरुष).

Example - तीर्थध्वाङ्क्षः 'a crow at the scared bathing place'.

Clue for the type identification - In a compound if the second component carries the word ध्वाङ्क्ष in the meaning of 'censure' then the

compound type may be identified as सप्तमीतत्पुरुष.

27. सूत्र - "क्तेनाहोरात्रावयवाः" (2-1-45)

Meaning - (a case-inflected word ending in the seventh case-termination signifying) parts of the day or night (is preferably compounded) with (a case-inflected word) derived with (the suffix) क्त (being called तत्पुरुष).

Examples - पूर्वाह्णकृतम् 'done in the morning', अपररात्रकृतम् 'done in the last part of the night'.

Clue for the type identification - In a compound, if the first component contains words relating with name of divisions of day or night and the second component contains a word with ending with क्त suffix the the compound type can be identified as सप्तमीतत्पुरुष.

28. सूत्र - "तत्र" (2-1-46)

Meaning - (the case-inflected word) तत्र "there" (ending in a seventh case-termination is preferably compounded with a case-inflected word derived by means of the suffix क्त (being called तत्पुरुष).

Example - तत्रभुक्तम् 'eaten there'.

Clue for the type identification - In a compound, if the first component contains the word तत्र and the second component contains the word which ends with क्त suffix then the compound type can be identified as सप्तमीतत्पुरुष.

29. सूत्र - "पात्रेसमितादयश्च" (2-1-48)

Meaning - also (the words) पात्रेसमिता etc (when the sense of abuse is to be conveyed, the compounds being called तत्पुरुष).

Examples - All these are irregular compounds. Some of the words contained in this list are compounds formed with the past-participle (क्त). Thus the following is the list of the words :- पात्रेसमिताः। पात्रेबहुलाः। उदुम्बरमशकाः। उदरकृमिः। कूपकच्छपः। कूपचूर्णकः। अवटकच्छपः। कूपमण्डूकः। कुम्भमण्डूकः। उदपानमःडूकः। नगरकाकः। नगरवायसः। मातरिषुरुषः। पिण्डीषूरः। पितरिषूरः। गेहेशूरः। गेहेनर्दी। गेहेक्ष्वेडी। गेहेविजिती। गेहेव्याडः। गेहेमेही। गेहेदाही। हेहेदृप्तः। गेहेधृष्टः। गर्भेतृप्तः। आखनिकबकः। गोष्ठेशूरः। गोष्ठे विजिती। गोष्ठेक्ष्वेडी। गोष्ठेपटुः। गोष्ठेपण्डितः। गोष्ठेप्रगल्भः। कर्णेटिट्टिभः। कर्णेटिरिटिरा। कर्णेचुरचुरा। The use of च in aphorism is to restrict it to these very words. Therefore we cannot compound परमाः पात्रेसमिताः, and it cannot be part of another compound.

30. सूत्र - "पूर्वकालैकसर्वजरत्पुराणनवकेवलाः समानाधिकरणेन" (2-1-49)

Meaning - (a case inflected word denoting an action) which precedes in time and (the case-inflected words) एक "single", सर्व "all", जरत् "new", पुराण "ancient", नव "new" and केवल "only" (only compounded) with a syntactically agreeing (case-inflected word, the compound being called तत्पुरुष).

Examples - स्नातानुलिप्तः "bathed and perfumed", कृष्टसमीकृतम् "ploughed and levelled", एकनाथः "havin one master", सर्वयाज्ञिकाः "all the members of a sacrifice", सर्वमनुष्याः "all men", जरन्नैयायिकाः "old logician", पुराणमीमांसकाः "old मीमासंकs" etc.

Clue for the type identification - The aphorism provides a list of

first components. In a compound, if first component contains one of these words then the compound type may be identified as a कर्मधारय.

31. सूत्र - "पापाणके कुत्सितैः" (2-1-54)

Meaning - (the case-inflected words) पाप "bad" and अणक "gossipy" (are compounded) with (syntactically agreeing, case-inflected words signifying) objects of contempt (the compound being called तत्पुरुष).

Description - Both the words पाप and अणक are words of contempts(कुत्सन) ; by the last aphorism, they have stood as second members in the compound ; the present aphorism, however is so framed with regard to aphorism 1-2-43 and 2-2-30 that they will stand as first.

Examples - पापनापितः or अणकनापितः 'a contemptible barber', पाप or अणक-कुलालः 'a contemptible potter'.

Clue for the type identification - The current aphorism explains a तत्पुरुष compound. In a compound, if the first component constains the words पाप or अणक then the compound type may be identified as a उभय-पद्-विशेषण-कर्मधारय. To implement this aphorism, one needs to prepare a list of कुत्सित words.

32. सूत्र - "उपमितं व्याघ्रादिभिः सामान्याप्रयोगे" (2-1-56)

Meaning - (a case-inflected word signifying) the object compared (is preferably compounded) with (the syntactically agreeing, case-inflected words) व्याघ्र "tiger", etc. provided that (a word signifying) the common property is not used (being called तत्पुरुष).

Description - This is a modification of last aphorism by which the substantive (उपमान) would have stood first, by the present, the उपमेय or the qualified stands first.

Example - पुरुषोऽयं व्याघ्र इव = पुरुषव्याघ्रः 'a person-tiger' (in strength). In previous aphorism उपमानानि सामान्यवचनैः" (2-1-55), the compounding was between the उपमान and the उपमेय. In the present, the compounding is between the उपमित and certain उपमान words but never with सामान्यवचन. In the current aphorism, the word व्याघ्रादिभिः suggests the व्याघ्रादिगण and the words belong व्याघ्रादिगण are :- 1) व्याघ्र "A tiger", 2) सिंह "A lion", 3) ऋक्ष "A bear", 4) ऋषभ "A bull", 5) चन्दन "Sandal", 6) वृक "A wolf", 7) वृष "A bull", 8) वराह "A boar, hog", 9) हस्तिन् "A elephant", 10) तरु "A tree", 11) कुञ्जर "A elephant", 12) रुरु "A kind of deer", 13) पृषत "The spotted antelope", 14) पुण्डरीक "A lotus flower", 15) पलाश "A tree, Butea Frondosa" and 16) कितव "A rogue, cheat".

Clue for the type identification - Any compound whose second component is from any of the above list, it is possibly an उपमान-उत्तरपद-कर्मधारय compound.

33. सूत्र - "विशेषणं विशेष्येण बहुलम्" (2-1-57)

Meaning - (a case-inflected word signifying) a qualifier (is) variously (compounded) with (a syntactically agreeing case-inflected word signifying) the item qualified (the compound conveying the same meaning as the formally corresponding wordgroup and being called तत्पुरुष).

Example - नीलोत्पलम् 'a blue lotus'.

Clue for the type identification - To implement this aphorism one needs a list of विशेषणs, which are typically marked in the dictionaries. Any compound whose first compound is विशेषण-पूर्वपद-कर्मधारय compound.

34. सूत्र - "पूर्वापरप्रथमचरमजघन्यसमानमध्यमध्यमवीराश्च" (2-1-58)

Meaning - (the case-inflected words) पूर्व "previous", अपर "later", प्रथम "first", चरम "last", जघन्य "low", समान "same", मध्य "middle", मध्यम "middle" and वीर "brave" (are preferably compounded with a syntactically agreeing case-inflected word and the compound being called तत्पुरुष).

Examples - पूर्ववैयाकरणः 'Previous grammarian', अपरवैयाकरणः 'later grammarian' etc.

Clue for the type identification - If the compound has any of the words from the list पूर्व, अपर, प्रथम, चरम etc. as its first components then it may be a विशेषण-पूर्वपद-कर्मधारय compound.

35. सूत्र - "श्रेण्यादयः कृतादिभिः" (2-1-59)

Meaning - (the case-inflected words) श्रेणि "guild" etc. (are compounded) with (the syntactically agreeing, case-inflected words) कृत "made", etc. (the compound being called तत्पुरुष).

Example - श्रेणीकृतम् 'made into guild' etc.

Clue for the type identification - If the first component of the compound is from श्रेण्यादिगण and the second component from the कृतादिगण then it is a कर्मधारय compound.

36. सूत्र - "क्तेन नञ्विशिष्टेनानञ्" (2-1-60)

Meaning - (a case-inflected word) not containig (the negative particle) नञ् (is preferably compounded) with (a syntactically agreeing case-inflected word ending in the suffix) क्त which is distinguished (from the other constituent) by (the fact that it does contain) नञ् (being called तत्पुरुष).

Examples - कृताकृतम् "done and not done" i.e. partly done and partly not done.

Clue for the type identification - If a compound has both the components ending in क्त and one as a negation of the other, then it is of type विशेषण-उभयपद-कर्मधारय compound.

37. सूत्र - "सन्महत्परमोत्तमोत्कृष्टाः पूज्यमानैः" (2-1-61)

Meaning - (the case-inflected words सत् "good", महत् "great", परम "highest", उत्तम "best" and उत्कृष्ट "eminent" (are preferably compounded) with (syntactically agreeing case-inflected words signifying) objects which are being honoured/praised (being called तत्पुरुष).

Description - सद्वैद्यः "a good physician". The महत् becomes महा in compound such as महापुरुषः "A great man" परमपुरुषः "the highest person", उत्तमपुरुषः "the best person", उत्कृष्टपुरुषः "the excellent person".

Clue for the type identification - If the first component contains adjectives words like सत्, महत्, परम, उत्तम and उत्कृष्ट then the compound is of type विशेषण-पूर्वपद-कर्मधारय compound.

38. सूत्र - "वृन्दारकनागकुञ्जरैः पूज्यमानम्" (2-1-62)

Meaning - (A case-inflected word signifying) an object which is being honoured/praised (is preferably compounded) with the syntactically agreeing, case-inflected words) वृन्दारक (a kind of) "deity", नाग "elephant" (being called तत्पुरुष).

Description - गोवृन्दारकः "An excellent bull or cow".

Clue for the type identification - It the second component contains adjectives words like वृन्दारक, नाग and कुञ्जर then the compound may be of type विशेषण-उत्तरपद-कर्मधारय compound.

39. सूत्र - "कतरकतमौ जातिपरिप्रश्ने" (2-1-63)

Meaning - (the case-inflected words) कतर "who (out of two)?", कतम "who (out of many)?" (are preferably compounded with a syntactically agreeing, case-inflected word, when a question (is asked) regarding (belonging to a particular) Vedic school (the compound being called तत्पुरुष).

Examples - कतरकठः "who (out of two) belongs to the कठ school", कतमकलापः "who (out of these many) belong to the कलाप school".

Clue for the type identification - If the first component contains the words कतर and कतम and the second component contains a जातिवाचक word then it is of type विशेषण-पूर्वपद-कर्मधारय.

40. सूत्र - "किं क्षेपे" (2-1-64)

Meaning - (the case-inflected word) किम् "who", which, what?" (is

compounded with a syntactically agreeing, case-inflected word) when (the compounded conveys the sense of) abuse (the compound being called तत्पुरुष).

Examples - किंराजा यो न रक्षति ? "he is a bad king who does not protect his subjects.", किं सखा योऽभिद्रुह्यति "he is a bad friend who hates".

Clue for the type identification - If the first component contains the word किम् then it is of type विशेषण-पूर्वपद-कर्मधारय compound.

41. सूत्र - "पोटायुवतिस्तोककतिपयगृष्टिधेनुवशावेहद्बष्कयणीप्रवक्तृश्रोत्रियाध्यापकधूर्तैर्जातिः" (2-1-65)

Meaning - (a case-inflected word signifying) a जाति "class/generic notion" (is compounded) with the syntactically agreeing, case-inflected words) पोटा "a female with male characteristics", युवति "a young female", स्तोक "a little", कतिपय "some", गृष्टि "a cow which had just one calf", धेनु "a sucking cow", वशा "a barren cow", वेहत् "a cow that miscarries", बष्कयणी "a cow having a young calf", प्रवक्तृ "one who expounds", श्रोत्रिय "a learned reciter of the Vedas", अध्यापक "a teacher of शास्त्रिक texts" and धूर्त "a shrewd one" (being called तत्पुरुष).

Examples - इभपोटा "a young female elephant", इभयुवतिः "a female elephant", अग्निस्तोकः "a little fire", उद्श्वत् कतिपयम् "a little butter milk", गोगृष्टिः, गोधेनुः, गोवशाः, गोवेहत्, गोबष्कयणी, कठप्रवक्ता "an expounder of कठ", कठश्रोत्रियः "a ब्राह्मण who has mastered the कठ branch of the यजुर्वेद", कठाध्यापकः "a teacher of the कठ branch of the यजुर्वेद".

Clue for the type identification - If the second component contains पोटा, युवति, स्तोक, कतिपय, गृष्टि, धेनु, वशा, वेहद्, बष्कयणी, प्रवक्तृ, श्रोत्रिय, अध्यापक

and धूर्त then it is of type विशेषण-उत्तरपद-कर्मधारय.

42. सूत्र - "प्रशंसावचनैश्च" (2-1-66)

Meaning - also (a case-inflected word signifying a जाति "class/generic notion" (is preferably compounded) with (syntactically agreeing case-inflected words) signifying praise (being called तत्पुरुष).

Description - गोमतल्लिका "an excellent cow", गोमचर्चिका "an excellent cow" etc.

Clue for the type identification - If the first component of the compound is a जाति and the second component is a word used in the sense of praising such as मतल्लिका, मचर्चिका, उद्ध, तल्लज etc.[4] then the compound may be of type विशेषण-उत्तरपद-कर्मधारय.

43. सूत्र - "युवा खलतिपलितवलिनजरतीभिः" (2-1-67)

Meaning - (the case-inflected word) युवन् "young" (is preferably compounded) with (the syntactically agreeing, case-inflected words) खलति "bald", पतित "grey-haired", वलिन "wrinkled" and जरति "old woman" (being called तत्पुरुष).

Description - युवखलतिः "bald in youth", युवपलितः "grey-headed in youth", युववलिनः "wrinkled in youth" and युवजरती "appearing old in youth" etc.

Clue for the type identification - In a compound, if the first component contains the word युवन् and the second component contains any one of the word from the list such as खलति, पलित, वलिन

[4]अमरकोषः

etc. then the compound type may be indentified as विशेषण-उत्तरपद-कर्मधारय.

44. सूत्र - "वर्णो वर्णेन" (2-1-69)

Meaning - (a case-inflected word signifying) a colour (is preferably compounded) with (a syntactically agreeing, case-inflected word signifying) a colour (being called तत्पुरुष).

Example - कृष्णसारङ्गः "spotted antelope".

Clue for the type identification - If both of the components of a compound are वर्णवाची then it is of type of विशेषण-उभयपद-कर्मधारय. To implement this aphorism one has to prepare the list of वर्णवाची words.

45. सूत्र - "चतुष्पादो गर्भिण्या" (2-1-71)

Meaning - (case-inflected words signifying) a quadruped (are preferably compounded) with (the syntactically agreeing, case-inflected word) गर्भिणी "pregnant" (being called तत्पुरुष).

Examples - गोगर्भिणी "a pregnant cow", अजागर्भिणी "a pregnant she-goat" etc.

Clue for the type identification - The current aphorism explains a तत्पुरुष compound where first component contains a word in the sense of quadruped and the second component contains the word गर्भिणी and the compound type can be indentified as a विशेषण-उत्तरपद-कर्मधारय compound. To implement this one needs a list of quadruped animals.

46. सूत्र - "मयूरव्यंसकादयश्च" (2-1-72)

Meaning - also (the compound) मयूरव्यंसक 'a cunnig peacock' etc are called तत्पुरुष.

Examples - All these are irregular compounds. Thus the following is the list of the words :- मयूरव्यंसकः । छात्रव्यंसकः । काम्बोजमुण्डः । यवनमुण्डः । छन्दसि हस्तेगृह्य । पादेगृह्य । लाङ्गलेगृह्य । पुनर्दाय । एहीडादयोऽन्यपदार्थे एहीडम् । एहियवं वर्तते । एहिवाणिजाक्रिया । अपेहिवाणिजा । प्रेहिवाणिजा । एहिस्वागता । अपोहिस्वागता । प्रेहिस्वागता । एहिद्वितीया । अपेहिद्वितीया । इहवितर्का । प्रोहकटा । अपोहकटा । प्रोहकर्दमा । अपोहकर्दमा । उद्धरचूडा । आहरचेला । आहरवसना । आहरवनिता । कृन्तविचक्षणा । उद्धरोत्सृजा । उद्धमविधमा । उत्पचिविपचा । उत्पतनिपता । उच्चावचम् । उच्चनीचम् । अचितोपचितम् । अवचितपराचितम् । निश्चप्रचम् । अकिञ्चनः । स्नात्वाकालकः । पीत्वास्थिरकः । भुक्त्वासुहितः । प्रोष्यपापीयान् । उत्पत्यपाकला । निपत्यरोहिणी । निषण्णाश्यामा । अपेहिप्रसवा । इहपञ्चमी । इहद्वितीया । जहिकर्मणा बहुलमाभीक्ष्ण्ये कर्तारं चाभिदधाति जहिजोडः । उज्जहिजोडः । जहिस्तम्बः । उज्जहिस्तम्बः । आख्यातमाख्यातेन क्रियासातत्ये अश्नीतपिबता । पचतभृज्जता । खादतमोदता । खादतवमता । खादताचमता । आहरनिवपा । आवपनिष्किरा । उत्पचच्चिपचा । भिन्धिलवना । छिन्धिविचक्षना । पचलवना । पचप्रकूटा । The use of च in aphorism is to restrict it to these very words.

47. सूत्र - "अर्धं नपुंसकम्" (2-2-2)

Meaning - (the case-inflected) neuter (word) अर्धम् "a half" (is preferably compounded with a case inflected word signifying a whole, provided that the whole mentioned is one single entity, the compound conveying the same meaning as the formally corresponding wordgroup and being called तत्पुरुष).

Example - अर्धपिप्पली "a half of the पिप्पली" etc.

48. सूत्र - "द्वितीयतृतीयचतुर्थतुर्याण्यन्यतरस्याम्" (2-2-3)

Meaning - (the case-inflected words) द्वितीय "second", तृतीय "third", चतुर्थ "fourth" and तुर्य "fourth" (are) indifferently (compounded with a case-inflected word signifying a whole, provided that the whole mentioned is one single entity (being called तत्पुरुष).

Example - द्वितीयभिक्षा "second begging" etc.

Reason - The clue exist in the first part of the component, viz. that it refers to signified portion, and thus is a पूरण-संख्या.

49. सूत्र - "नञ्" (2-2-6)

Meaning - (the case-inflected words) नञ् (is preferably compounded with a case-inflected word (being called तत्पुरुष).

Examples - अब्राह्मणः "not a Brahmana or Non-Brahmana", अनश्वः "Not a horse" etc.

Clue for the type identification - If the first component contains the word 'अ' or 'अन्' then the compound type may be indentified as नञ्-तत्पुरुष compound.

50. सूत्र - "ईषदकृता" (2-2-7)

Meaning - (the case-inflected word) ईषत् "to some extent, slightly" (is preferably compounded with a case-inflected word) not (ending in a कृत्(-suffix)(being called तत्पुरुष).

Example - ईषत्पिङ्गलः "a little brownish".

Reason - If the first component of the compound is ईषत् then it is likely that the compound is a तत्पुरुष. However for confirmation,

one needs to know whether the word ईषत् is used in the sense of slightly/to some extent. This can be done only manually.

51. सूत्र - "याजकादिभिश्च" (2-2-9)

Meaning - also (a case-inflected word ending in the sixth case termination is preferably compounded) with (the case-inflected words) याजक "one who worships/pays honour to" etc. (being called तत्पुरुष).

Examples - ब्राह्मणयाजकः "a Brahmana's sacrificer" and the words which comes under the याजकादिगण are :- 1) याजक, 2) पूजक, 3) परिचारक, 4) परिवेषक, 5) परिषेचक, 6) स्नापक or स्नातक, 7) अध्यापक, 8) उत्साहक or उत्सादक, 9) उद्वर्तक, 10) होतृ, 11) भर्तृ, 12) रथगणक, 13) पत्तिगणक, 14) पोतृ, 15) हर्तृ and 16) वर्तक.

Reason - The current aphorism explains a तत्पुरुष compound and provides a list of some irregular compounds. We treat these as special compounds and do not anlyse them further.

52. सूत्र - "कुगतिप्रादयः" (2-2-18)

Meaning - (the case-inflected word) कु "evil", (the case inflected words called) गति and (the case-inflected words mentioned as) प्रादयः "प्र,परा,अप... etc." (are always compounded with a case-inflected word the compound, being called तत्पुरुष).

Description - कुपुरुषः "a sinful man", उररीकृतम् "having asserted", दुष्पुरुषः "a bad man".

The current aphorism explains a तत्पुरुष compound and provides us

three types of compounds 1) कुसमास, 2) गतिसमास and 3) प्रादिसमास.

(a) कुसमास is used in the sense of निन्दा "censure". For example कुपतिः "a bad husband". By looking the first component the type of this kind of compound can be identified easily and can be implemented.

(b) गतिसमास is found in many senses and need so many informations for implementation such as the second compound always should be a क्रिया not the तिङ्, first component may contain ऊर्यादिगण words or ending with the च्वि or डाच् suffix etc. For examples उररीकृत्य, शुक्लीकृत्य, पटपटाकृत्य etc.

(c) प्रादिसमास is found in many senses and can be implemented easily for type identification because the first component alwayse contains a prefix प्र, परा, अति etc. and the second component will be a case-inflected word. For instance प्राचार्यः.

6.2.3 बहुव्रीहि

53. सूत्र - "संख्ययाव्ययासन्नदूराधिकसंख्याः संख्येये" (2-2-25)

Meaning - (case-inflected words called) अव्यय "indeclinable", (the case-inflected words) आसन्न "near", अदूर "not far", अधिक "additional", and (the case-inflected words called) संख्या "numeral" (are compound) with (a case-inflected word called संख्या "numeral", when (the compound) conveys the sense of संख्येय "item to be counted" (the compound being called बहुव्रीहि).

Examples - उपदशाः "those who are neat to ten" or उपविंशः "nineteen or twenty-one".

Clue for the type identification - If the first component contains the indeclinable words and the words like आसन्न, अदूर, अधिक and second component contains the words in sense of the numeral then the compound type may be identified as संख्योत्तरपद-व्यधिकरण-बहुव्रीहि compound.

54. सूत्र - "दिङ्मामान्यन्तराले" (2-2-26)

Meaning - (case-inflected words which are) names of directions (are invariably compounded with each other, the compound) conveying the sense of अन्तराल "intermediate region" and being called बहुव्रीहि.

Examples - दाक्षणपूर्वादिक् "the direction midway between southe and east", पूर्वोत्तरा "north-east"

Clue for the type identification - To implement this rules, one needs to prepare the list of दिङ्मा "names of directions" and has to see whether both of the components contain the words from दिङ्मा list then the compound type may be identified as दिग्वाचक-बहुव्रीहि compound.

55. सूत्र - "तेन सहेति तुल्ययोगे" (2-2-28)

Meaning - (the case-inflected word) सह "together with" (is always compounded with a case-inflected word represented by the word-pattern) तेन (the compound) conveying the sense of तुल्ययोग "equal connection" (with an action), and being called बहुव्रीहि.

Example - सपुत्रः "with son" etc.

Clue for the type identification - If the first component contains

the word स then it is possible that the compound is of type सह-पूर्वपद-व्यधिकरण-बहुव्रीहि. In case the gender of the second component is different from its default gender, then the compound is definitely of this type. However, if the gender of the second component is same as its regular gender, than there is also a possibility of this compound being an अव्ययीभाव[5].

Below we summarise the conditions and extra semanitc information needed for automatic detection of compund type.

		Condtions			
S.No.	Sūtra Reffr.	Initial word	Final word	Extra-Info	Type
1	2-1-6	All the prefixs or indeclinables like प्र, परा, अप, यथा etc.	neuter gender	-	A1
2	2-1-8	यावत्	Neuter gender	-	A1
3	2-1-9	सु, प्रति	Neuter gender	-	A1
4	2-1-10	अक्ष, शलाका or संख्याः like एक, द्वि etc.	परि	-	A2
5	2-1-12	अप, परि, बहिस् or word endig with अञ्च्	Neuter gender or fifth-case ending word	-	A1

[5]पा०सू० - अव्ययं विभक्तिसमीप... "2-1-6"

6	2-1-13	आङ्	Neuter gender or fifth-case ending word	-	A1
7	2-1-14	अभि, प्रति	Neuter gender or fifth-case ending word	-	A1
8	2-1-15	अनु	Neuter gender	-	A1
9	2-1-16	अनु	Neuter gender	-	A1
10	2-1-17	तिष्ठद्गुप्रभृतीनि च	तिष्ठद्गु etc.	-	A3
11	2-1-18	पारे, मध्ये	fifth-case affix ending word	-	A7
12	2-1-19	The numerals like एक, द्वि	वंश्यवाची words	वंश्यवाची list	A6
13	2-1-20	The numerals like एक, द्वि	नदीवाची words	नदीवाची list	A6
14	2-1-24	-	श्रित, अतीत, पतित, गत, अत्यस्त, प्राप्त, आपन्न	-	T2
15	2-1-25	स्वयम् or variation of स्वयम्	क्तान्त word	-	T2
16	2-1-26	खट्वा	क्तान्त word	-	T2
17	2-1-27	सामि	क्तान्त word	-	T2

18	2-1-28	the list of the words which denote the time	क्तान्त word	list of words denoting time	T2
19	2-1-31	-	पूर्व, सदृश, सम, ऊ-नार्थ, कलह, निपुण, मिश्र, श्लक्ष्ण	-	T3
20	2-1-36	-	तदर्थ, अर्थ, बलि, हित, सुख, रक्षित	-	T4
21	2-1-37	-	भय, भीति, भी	-	T5
22	2-1-38	-	अपेत, अपोढ, मुक्त, पतित, अपत्र	-	T5
23	2-1-39	स्तोका,अन्तिक, दूरार्थ, कृच्छ्र	क्तान्त word	-	T5
24	2-1-40	सप्तमी शौण्डैः	अक्षशौण्डः	-	T7
25	2-1-41	-	सिद्ध, शुष्क, पक्व, बन्ध	-	T7
26	2-1-42	-	ध्वाङ्क्ष	-	T7
27	2-1-45	The words relating with name of divisions of day or night	क्तान्त word	list of words denoting names of divisions of days and night	T7

28	2-1-46	तत्र	क्तान्त word	-	T7
29	2-1-48	पात्रेसमितादयश्च	पात्रेसमिताः, पात्रेब-हुलाः etc.	-	T7
30	2-1-49	पूर्व, काल, एक, सर्व, जरत, पुराण, नव, केवल	-	-	K1
31	2-1-54	पाप, अणक	word related to कुत्सित	list of कुत्सित words	K3
32	2-1-56	उपमितं व्याघ्रादिभिः सामान्याप्रयोगे	पुरुषव्याघ्रः etc.	-	K5
33	2-1-57	any विशेषण	-	list of विशेषणs	K1
34	2-1-58	पूर्व, अपर, प्रथम, चरम, जघन्य, समान,मध्य, मध्यम, वीर	-	-	K1
35	2-1-59	श्रेण्यादयः कृतादिभिः	श्रेणीकृतम् etc.	-	K
36	2-1-60	क्तान्त word	क्तान्त word with negation अ	-	K2
37	2-1-61	सत, महत, परम, उत्तम, उत्कृष्ट	-	-	K1
38	2-1-62	-	वृन्दारक,नाग, कुञ्जर	-	K2
39	2-1-63	-	कतर, कतम	-	K1
40	2-1-64	किम्	-	-	K1

41	2-1-65	जातिवाचक word	पोटा, युवति, स्तोक, कतिपय, गृष्टि, धेनु, वशा, वेहत्, बष्कयणी, प्रवक्तृ, श्रोत्रिय, अध्यापक, धूर्त	list of जातिवाचक words	K2
42	2-1-66	जातिवाचक word	words signifying praise such as मतल्लिका, मचर्चिका, उद्ध, तल्लज etc.	list of जातिवाचक and प्रंशसावाचक words	K2
43	2-1-67	युवन्	खलति, पलित, वलिन, जरति	-	K2
44	2-1-69	वर्णवाचक word	वर्णवाचक word	list of वर्णवाचक words	K3
45	2-1-71	चतुष्पादवाचक word	गर्भिणी	list of वर्णवाचक words	K2
46	2-1-72	मयूरव्यंसकादि	मयूरव्यंसकः etc.	-	K
47	2-2-2	अर्धं नपुंसकम्	अर्धपिप्पली	-	T1
48	2-2-3	द्वितीयतृतीयचतुर्थ...	द्वितीयभिक्षा etc.	-	K
49	2-2-6	अ/अन्	-	-	Tn
50	2-2-7	ईषत्	-	-	T
51	2-2-9	याजकादिभिश्च	ब्राह्मणयाजकः etc.	-	K
52	2-2-18	कु	-	-	Tk

52.1	2-2-18	प्रादय - प्र, परा, निस्, निर् etc.	-	-	Tg/Tp
53	2-2-25	अव्ययs, आसन्न, दूर, अधिक, संख्या	संख्या in the sense of संख्येय like दश, विंश	list of संख्यावाचक words with special gender[6]	Bvs
54	2-2-26	names of directions	names of directions	list of दिङ्नाम "names of directions"	Bsd
55	2-2-28	स	words with special gender	-	BvS
55.1	2-2-28	स	words	-	A1

Till now we have explained the aphorisms which we have implemented to identify the types of various compounds. Now we are going to explain the aphorisms which we could not implement and the reason(s) thereof.

1. सूत्र - "यथाऽसादृश्ये" (2-1-7)

   Meaning - (the indeclinable) यथा (is invariably compounded with a case-inflected word) in a meaning other than सादृश्य "similarity" (the compound being called अव्ययीभाव).

[6]The second components in बहुव्रीहि compounds contain the special gender like in द्वित्राः, the stem त्रि contains the mesculine gender त्राः etc.

Example - यथावृद्धम् = ये ये वृद्धाः (every old person).

Reason - The current aphorism explains an अव्ययीभाव compound and its an exceptional rule of 2-1-6. Usally the word यथा is used in four types of meanings :- 1) योग्यता "ability", 2) वीप्सा 3)पदार्थानतिवृत्ति and 4) सादृश्य "similarity". Except the fourth one सादृश्य, one of the meaning of rest of three should be there in the compound then only the compound may be identified as an अव्ययीभाव. It is not possible to check whether the meaning is in the desired sense or not without doing any semanitc analysis of the context. Hence it is not possible to decide the type of the given compound.

2. सूत्र - "अत्यन्तसंयोगे च" (2-1-29)

Meaning - also (case-inflected words ending in the second case-termination, signifying a period of time, are compounded with a case-inflected word) when (the sense of) अत्यन्तसंयोग "invariable, uninterupted connection" is to be conveyed (the compound being called तत्पुरुष).

Example - मुहूर्तसुखम् "a momentary pleasure".

Reason - It is not possible to decide just by looking at the word whether there is अत्यन्तसंयोग or not and hence it is not possible to classify such compounds automatically.

3. सूत्र - "तृतीया तत्कृतार्थेन गुणवचनेन" (2-1-30)

Meaning - (a case-inflected word) ending in the third case-termination (is preferably compounded) with (a case-inflected word) standing for a quality which is an entity caused/brought

about by (what is reffered to) by that (being called तत्पुरुष).

Examples - शङ्कुलाखण्डः "cut by nipper", धान्यार्थः "wealth acquired by grain".

Reason - To classify such compounds, it is neccessary to decide semantic relation between the components, and this requires the knowledge of योग्यता - semantic compatibility between the word meanings. Therefore it is not possible to classify such compounds purely on the basis of syntax or lexicon.

4. सूत्र - "कर्तृकरणे कृता बहुलम्" (2-1-32)

   Meaning - (a case-inflected word ending in the third case-termination) which conveys the sense of कर्तृ "agent" or करण "instrument" (is) variously (compounded) with (a case-inflected word) ending in a कृत्-suffix (being called तत्पुरुष).

   Examples - अहिना हतः = अहिहतः "killed by a snake", नखभिन्नः "divided by the nails" etc.

   Reason - To implement this aphorism, the knowledge of योग्यता is required.

5. सूत्र - "कृत्यैरधिकार्थवचने" (2-1-33)

   Meaning - (a case-inflected word ending in the third case-termination conveying the sense of कर्तृ "agent" or of करण "instrument" is compounded) with (a case-inflected word) derived by means of a कृत्य(-suffix), provided that (the compound) conveys the sense of exaggeration (being called तत्पुरुष).

Examples - वातच्छेद्यं तृणं 'thin grass' (so fragile that it can be cut by wind) in the sense of praise, denoting softness, or it may be the reverse ; denoting weakness.) So also काकपेया नदी 'a full river' (so full that a crow may dip his beak into it, and drink while sitting on the bank. In this sense it is praise. It may be censure also, and it will mean a shallow river that a crow may dip his beak into it and touch the bottom and drink.)

Reason - Since this aphorism relies heavily on the context, and the semantics it is not possible to implement this rule computationally.

6. सूत्र - "अन्नेन व्यञ्जनम्" (2-1-34)

Meaning - (a case-inflected word ending in the third case-termination) signifying a flavouring ingredient (is preferably compounded) with (a case-inflected signifying) an article of food (being called तत्पुरुष).

Description - That which is to be prepared is called अन्न and that which prepares is व्यञ्जन as दध्योदनः 'rice perpared or made relishable with curd.' The words 'food' and 'condiment' as represented in the above compounds, are connected in the sense by a verb understood.

Reason - To identify such compounds as तृतीयातत्पुरुष, one has to provide a list of words of अन्नवाची and corresponding व्यञ्जनवाची.

7. सूत्र - "भक्ष्येण मिश्रीकरणम्" (2-1-35)

Meaning - (a case-inflected word ending in the third case-termination signifying) an admixture (is preferably compounds)

with (a case-inflected word signifying) something edible (being called तत्पुरुष).

Example - गुडधाना 'barley prepared with juggery'.

Anything eatable, whether hard or soft, is called भक्ष्य, its refinement is called मिश्रिकरण.

Reason - As in the previous aphorism, to identify these type of compounds a तृतीयातत्पुरुष, one needs a lexicon of भक्ष्य and means of मिश्रिकरण for the corresponding भक्ष्यs. In the absence of such lexicon, it is not possible to identify such compounds.

8. सूत्र - "कृत्यैर्ऋणे" (2-1-43)

Meaning - (a case-inflected word ending in the seventh case-termination is compounded) with (a case-inflected word) ending in a कृत्य-suffix, when (the compunded conveys) the sense of debt (the compunded being called तत्पुरुष).

Example - मासदेयम् "a debt repayble within a month".

The word ऋण indicates by implication any appointed time in general, and not merely a time for the payment of a debt. Therefore we get compounds in the following case also - पूर्वाह्णगेयं साम "the साम that would be sung in the morning".

Reason - Since this involves semantic interpretation, it is not easy to detect the type of such compounds automatically.

9. सूत्र - "संज्ञायाम्" (2-1-44)

Meaning - (a case-inflected word ending in the seventh case-

termination is compounded with a case-inflected word) provided that (the compound formed) is a name (the compound being called तत्पुरुष).

Description - A संज्ञा is expressed by the complete word, hence it is an invariable (nitya) compound ; for we cannot express an appellative by a sentence. For instance अरण्येतिलकाः "wild sesamum" yielding no oil ; anything which does not answer to one's expectation.

Reason - This aphorism also needs the knowledge of the real world to interpret the relation between the words involved and thus is out of scope of mechanical identification.

10. सूत्र - "क्षेपे" (2-1-47)

Meaning - (a case-inflected word ending in the seventh case-termination is compounded with a case-inflected word ending in the suffix क्त) when (the compound conveys) the sense of abuse (the compound being called तत्पुरुष).

Example - अवतप्तेनकुलस्थितं त एतत् "thy this work is as if an ichneumon (नकुल) standing on hot ground" etc.

Reason - Since this involves semantic interpretation, it is not easy to detect the type of such compounds automatically.

11. सूत्र - "दिक्संख्ये संज्ञायाम्" (2-1-50)

Meaning - (a case-inflected word signifying) a direction or a number (is compounded with a syntactically agreeing case-inflected word) provided that (the compound formed is) a name (the compound

being called तत्पुरुष).

Examples - पूर्वेषुकामशमी "the town of इषुकामशमी in the East", सप्तर्षयः "seven sages" (the constellation of Great Bear) etc.

Reason - Since this involves semantic interpretation, it is not easy to detect the type of such compounds automatically.

12. सूत्र - "तद्धितार्थोत्तरपदसमाहारे च" (2-1-51)

Meaning - also (a case-inflected word signifying a direction or a number is compounded with a syntactically agreeing case-inflected word) provided that (the form derived) conveys the meaning of a तद्धित(-suffix), or that a final compound-constituent follows, or that (the form derived) conveys the sense of समाहार "group" (the compound being called तत्पुरुष).

Examples - पौर्वशालः, आपरशालः, अपरशालप्रियः etc.

The current aphorism explains a तत्पुरुष compound. To implement this aphorism, it needs the whole तद्धित environment and also the context.

13. सूत्र - "संख्यापूर्वो द्विगुः" (2-1-52)

Meaning - (a compound conditioned by 2-1-51) whose first constituent is (a word signifying) number (is called तत्पुरुष and) द्विगु.

Description - In a case where the sense is that of a तद्धित affix or when an additional member comes after the compound (उत्तरपदे) or when an aggregate (समाहारे) is to be expressed, the compound, the first member of which is numeral, is called द्विगु.

Example - पञ्चगवम् "an aggregate of five cows".

Reason - The anuvritti of तद्धित and समाहारे is coming to this aphorism. To implement it, we need तद्धित environment and the context of aggregate. It is not possible to decide type mechanically.

14. सूत्र - "कुत्सितानि कुत्सनैः" (2-1-53)

Meaning - (case-inflected words signifying) objects of contempt (are compounded) with (syntactically agreeing case-inflected words signifying) abuse (the compound being called तत्पुरुष).

Examples - वैयाकरणखसूचिः 'A bad grammarian, who does not know grammar, but contemplates the heaven (ख) when asked any questions, मीमासंकदुर्दुरूढः 'an evil conducted मीमांसक'.

Reason - In a compound, if the first component contains an adjective and the second component contains also an adjective in sense of contempt then the compound type can be identified as a उभयपदविशेषण-कर्मधारय compound. Since it not easy to decide whether the second component such as खसूचिः is used in the sense of abuse or not, it is not possible to mark such compound types automatically.

15. सूत्र - "उपमानानि सामान्यवचनैः" (2-1-55)

Meaning - (case-inflected words signifying) the standard of comparison (are compounded) with (syntactically agreeing, case-inflected words) signifying a common property (the compound being called तत्पुरुष).

Description - That by or to which a thing is compared is 'उपमान' and

the उपमेय (the thing compared) is called सामान्य or common. Thus घन-श्यामः कृष्णः 'cloud-black Krishna' (Krishna black as a cloud). Here श्याम is a quality common to Krishna and cloud, therefor 'cloud' which is the उपमान is compounded with it. So also कुमुदश्येनी 'lily-white', हंसगद्गदा 'Swan sounding', न्यग्रोध-पारिमण्डला 'globular as न्यग्रोध tree'.

In anlaysing the above compound घनश्यामः we must use the word इव, as घन इव श्यामः, and this shows that it is merely a simile or metaphorical use of the word. This aphorism is made in order to declare a niyama rule, so that the उपमान word should stand first.

Reason - To implement this aphorism, one needs a list of उपमानवाची words and the corresponding उपमेयवाची words.

16. सूत्र - "कृत्यतुल्याख्या अजात्या" (2-1-68)

Meaning - (case-inflected words ending in a कृत्य(-suffix) and synonyms of तुल्य "equal" (are preferably compounded) with (a syntactically agreeing case-inflected word signifying) something other than a जाति "class/generic notion" (being called तत्पुरुष).

Examples - भोज्योष्णम् "hot food", तुल्यश्वेतः "equally white", सदृशश्वेतः "equally white" etc.

Reason - To implement this aphorism, the second component should not refer to a class. This is possible only if the lexicon is rich with this kind of information.

17. सूत्र - "पूर्वापराधरोत्तरमेकदेशिनैकाधिकरणे" (2-2-1)

Meaning - (the case-inflected words) पूर्व "fore (part)", अपर "hind

(part)", अधर "lower (part)" and उत्तर "upper (part)" are preferably compounded with (a case inflected word) signifying (a whole) having parts, provided that (the whole mentioned) is one single item, (being called तत्पुरुष).

Examples - पूर्वकायः "the front of the body", अपरकायः "the black of the body".

Reason - The clue viz. second word indicating it refers to a body with parts is semantic. So unless a lexicon rich with such an information is available, such compound types can't be identified mechanically.

18. सूत्र - "नित्यं क्रीडाजीविकयोः" (2-2-17)

Meaning - (a case-inflected word ending in the sixth case-termination, when the genitive case ending conveys the sense of कर्मन् "direct object" is) always compounded with a case-inflected word ending in the suffixes तृच् or अक) when (the compound) conveys the sense of क्रीडा "a game" or जीविका "a profession" (the compound being called तत्पुरुष).

Examples - उद्दालकपुष्पभञ्जिका "a sort of game played by the people in the eastern districts in which उद्दालक flowers are broken or crushed", दन्तलेखकः "one who earns his bread by painting or marking the teeth", नखलेखकः "a nail-painter by profession" etc.

Reason - The current aphorism explains a तत्पुरुष compound. To implement this rule one needs a list of words indicating either a क्रीडा or जीविका.

19. सूत्र - "उपपदमतिङ्" (2-2-19)

    Meaning - an उपपद which (does) not (end in a suffix called) तिङ् (is always compounded the compound being called तत्पुरुष).

    Examples - कुम्भकारः "one who makes pots", नगरकारः "one who makes the city" etc.

    Reason - If the second component is a kind of कृदन्त word which is found only in compounds such as कार, ज्ञः, गः etc. then the compound is likely to be उपपद-तत्पुरुष compound. For this we need a special morph module that handles such forms.

20. सूत्र - "अनेकमन्यपदार्थे" (2-2-24)

    Meaning - two or more (case-inflected words are invariable compounded) when (the compound) conveys the sense of a word other (than the compound-constituents, the compound being called बहुव्रीहि).

    Examples - प्राप्तमुदको ग्रामः "a water reached village", वीरपुरुषकः ग्रामः "a village possessed of heroic men" etc.

    Reason - There is no clue whatsoever except the fact that this word refers to something which is different from the refrents of the components involved. Thus it needs extra Linguisitc knowledge to identify the type of such compounds.

21. सूत्र - "चार्थे द्वन्द्वः" (2-2-29)

    Meaning - (two or more case-inflected words are invariably

compounded) when (the compound) conveys the sense of च "and" (the compound being called) द्वन्द्व.

Examples - रामलक्ष्मणौ "Rama and Lakshman", धवखदिरौ "the Mimosea and the Grislea" etc.

Reason - Just by looking at the components, the type-identification of this kind of compounds, is very difficult because the decision involves the meaning of components.

Finally we discuss the rules which deal with special or exceptional cases of compounds, which can not be analysed further. We just list them as exceptional compounds. Here, we are giving the list of exceptional or special rules.

1. सूत्र - "अन्यपदार्थे च संज्ञायाम्" (2-1-21)

   Meaning - Also, (a case-inflected word is compounded with a case-inflected word signifying a river) when the sense of अन्यपदार्थ an additional meaning (not expressed by the compound constituents) is to be conveyed, provided the (the compound formed) is a name (the compound being called अव्ययीभाव).

   Description - This is a विधिसूत्र prescribing compound formation. A word ending in a case affix is compounded with words denoting the names of rivers, when the compound word denotes a thing other than that expressed by the terms of the compound, and is an appellative ; the compound so formed being an अव्ययीभाव.

   The anuvritti of the word संख्या does not extend to this aphorism. Though this rule is given in the subdivision relating to optional compounds, it is, however, a nitya-samāsa rule : for no name (संज्ञा)

can ever be expressed by a sentence, and that being so, these compounds can never be analysed. As उन्मत्तगङ्गम् 'the country called उन्मत्तगङ्गम्.'

Since such compounds can never be analysed, they are handled as special compounds, and are termed as नद्युत्तरपद-अन्यपदार्थे-अव्ययीभाव.

2. सूत्र - "कुमारः श्रमणादिभिः" (2-1-70)

   Meaning - (the case-inflected word) कुमार "boy" (is preferably compounded) with (the syntactically agreeing, case-inflected words) श्रमणा "Buddhist/Jaina nun", etc. (being called तत्पुरुष).

   Example - कुमारीश्रमणा or कुमारश्रमणा "a virgin ascetic".

3. सूत्र - "प्राप्तापन्ने च द्वितीयया" (2-2-4)

   Meaning - also (the case-inflected words) प्राप्त "who/which has reached/obtained" and आपन्न "who/which has obtained" (are preferably compounded with a case-inflected word ending in the second case termination (being called तत्पुरुष).

   Examples - प्राप्तजीविकः (प्राप्तो जीविकां) as well as जीविकाप्राप्तः 'obtained his livelihood', आपन्नजीविकः and जीविकापन्नः.

   Reason - This is a special compound with only two possible words as second component. These are just treated as special cases and not analysed further.

4. सूत्र - "कालाः परिमाणिना" (2-2-5)

   Meaning - (case-inflected words signifying a period of) time (are

preferably compounded) with (a case-inflected word signifying) an item to be measured (being called तत्पुरुष).

Examples - मासो जातस्य = मासजातः "a month old", संवत्सरजातः "a year old" etc.

The current aphorism explains a तत्पुरुष compound and it needs the information (of परिमाण) the words denoting the object whose duration is measured by the time. We treat it as exceptional rule.

5. सूत्र - "अमैवाव्ययेन" (2-2-20)

Meaning - (a case-inflected word called उपपद is always compounded) with (a case-inflected word called) अव्यय "indeclinable" (ending in the suffix) अम् "only" (the compound being called तत्पुरुष).

Examples - स्वादुङ्कारं भुङ्क्ते "he eats having made his food sweet", लवणंकारम् "having seasoned" etc.

Reason - If the component is a word ending in second case suffix and the second component is a कृदन्त used in उपपद-तत्पुरुष compounds, then the compound may be classified under उपपद-तत्पुरुष.

6. सूत्र - "तृतीयाप्रभृतीन्यन्यतरस्याम्" (2-2-21)

Meaning - (the case-inflected words called उपपद) starting from (those prescribed as ending in) the third case termination (are) indifferently (compounded with a case-inflected word called अव्यय "indeclinable", ending in the suffix अम् only (being called तत्पुरुष).

Example - मूलकोपदंशं भुङ्क्ते or मूलकेनोपदंशं भुङ्क्ते "he eats after having relished the food with radish".

Type of compound -

Reason - The current aphorism explains a उपपद-तत्पुरुष compound. These type of compounds can be analysed provided कृदन्त morpholgical analyser analyses special forms. In the absence of such morpholgical analyser, we treat all such compounds as special cases.

7. सूत्र - "तत्र तेनेदमिति सरूपे" (2-2-27)

Meaning - two (case-inflected words) alike in form (and represented by the word-pattern) तत्र (or) तेन (are compounded with each other, the compound) conveying the meaning इदम् "this" (and being called बहुव्रीहि).

Examples - केशाकेशि "hair to hair, fighting by pulling each other's hair", दण्डादण्डी "stick against stick, fight with and stave" etc.

Reason - The current aphorism explains a प्रहरण-विषयक-बहुव्रीहि compound. These type of compounds can be analysed provided the morpholgical analyser analyses special forms such as केशा and केशि etc. In the absence of such morpholgical analyser, we treat all such compounds as special cases.

List of exceptional or special rules				
No.	Sūtra reff.	Sūtra	Example	Type
1	2-1-21	अन्यपदार्थे च संज्ञायाम्	उन्मत्तगङ्गम्	A5

2	2-1-70	कुमारः श्रमणादिभिः	कूमारीश्रमणा, कु-मारश्रमणा etc.	K
3	2-2-4	प्राप्तापन्ने च द्वितीयया	प्राप्तजीविकाः etc.	K
4	2-2-5	कालाः परिमाणिना	मासजातः, संवत्सरजातः etc.	K
5	2-2-20	अमैवाव्ययेन	स्वादुङ्कारं भुङ्क्ते।	-
6	2-2-21	तृतीयाप्रभृतीन्यन्यतरस्याम्	मुलकोपदंशं भुङ्क्ते।	-
7	2-2-27	तत्र तेनेदमिति सरूपे	केशाकेशि, दण्डा-दण्डी	Bsp

6.3 Statistical Apporach

In this section we describe the statistical approach. We use manually tagged corpus as a model for predicting the tags, given a pair of constituents of a compound. Manually tagged corpus has been generated by the SHMT consortium funded by DIT. The total corpus is around 600K words of which 80,155K words are compounds. These texts were tagged using the tagset given in appendix-B. All these compounds are thus tagged `in context' and contain only one tag. This corpus formed the training data. Another small corpus with 400 tagged compounds from totally different texts, was kept aside for testing.

We observe that certain words have affinity for certain other words. For example घन has an affinity towards a word indicating the colour such as

श्याम or name of the colour has an affinity towards an object in the real world such as flower as in नीलोत्पलम् etc. Further from the manually tagged data we also know the likely tag for such compounds. The manually tagged data also reveals that certain words when used as the first component of a compound indicate a particular compound. For example the indeclinables such as यथा as the first component always indicates an अव्ययीभाव, which is also supported by the Pāṇini's rule (अव्ययं विभक्तिसमीपसमृद्धिव्यृद्ध्यर्थाभावात्यया-सम्प्रति... "2-1-6").

6.3.1 Some features of the manually tagged data

1. Around 25% of the compounds were repeated.
2. The 38,594 tokens of compounds contain 14,203 types of left word and 24,391 types of right word.
3. The frequency distribution of highly frequent tags is shown in Table 6.10. To study the effect of fine-grained-ness we also merged the sub-types. Table 6.12 gives the frequency distribution of major tags, after merging the sub-types.

Coarse and fine grain tagset

Pāṇini has classified the compounds in four major classes however, for the purpose of generating a paraphrase, these four class are not sufficient. Hence fine grain tagset of 55 tags was evolved by the Sanskrit consortium. The performance of any classifier is inversely proportional to the number of tags. More the tags, less is the accuracy of a word getting correct tag. This is true with the compound type identification as well. Especially because of the associated words and their frequencies. There are more chances of the

data becoming sparse. Hence we decided to measure the performance of our type identifier using both the fine grained as well as coarse grained tagset. For the purpose of identification, we define our coarse grained tagset to be consisting of eight tags viz. तत्पुरुष, कर्मधारय, बहुव्रीहि, द्वन्द्व, उपपद, अव्ययीभाव.

Tag	% of words
T6	29.12
K1	12.05
Bs6	6.97
T3	5.64
Tn	4.50
Di	3.75
U	3.64

Table 6.4 : Distribution of Fine-grain-Tags

Tag	% of words
T	50.51
K	19.01
B	12.53
D	4.70
U	3.66
A	0.94
S	0.78

Table 6.6 : Distribution of Coarse-Tags

From the manually tagged data, we define the following probabilities

(a) P(T/w_1-w_2) = probability that the compound w_1-w_2 has a tag T.

$$= \frac{\#(w_1 - w_2 \; of \; type \; T)}{\#(w_1 - w_2)}$$

(b) P(T/w_1-) = probability that a compound with w_1 as the initial component has tag T.

$$= \frac{\#of \; words \; with \; w_1 \; as \; initial \; component \; with \; tag \; T}{\#of \; words \; with \; w_1 \; as \; an \; initial \; component}$$

(c) P(T/-w_2) = probability that a compound with w_2 as the final

component has tag T.

$$= \frac{\#(of\ words\ with\ w_2\ as\ initial\ component\ with\ tag\ T)}{\#of\ words\ with\ w_2\ as\ an\ initial\ component}$$

6.3.2 Algorithm

Given w_1-w_2

(i) If $P(T/w_1\text{-}w_2)$ exists, we choose the max $P(T_i/w_1\text{-}w_2)$

(ii) else if $P(T_i/w_1\text{-})$ and $P(T_i/\text{-}w_2)$ exist, we choose maximum $P(T_i/w_1\text{-}) \times P(T_i/\text{-}w_2)$

(iii) else if only $P(T_i/w_1\text{-})$ exists, we choose maximum $P(T_i/w_1\text{-})$

(iv) if only $P(T_i/\text{-}w_1)$ exists, we choose maximum $P(T_i/\text{-}w_1)$

(v) and finally if both w_1- and -w_2 do not occur in the manually tagged corpus, we assign a tag with maximum probability.

6.3.3 Performance Evaluation

The test data of 400 words are tagged `in context'. While our compound tagger does not see the context, and thus suggests more than one possible tag, and ranks them. Normally a tool is evaluated for its coverage and precision. In our case, the tool always produces tags with weights associated with them. Hence, the coverage is 100%. The precision is evaluated based on the ranks of the correct tags. Table6.14 shows results of 400 words with a coarse as well as fine grained tagset.

Thus, if we consider only the first rank, the precision with 8 tags is 76.1% and with 55 tags, it is 62.0%. The precision increases substantially to 95.8% and 83.5% respectively if we take first three ranks into account.

	with 55 tags		with 8 tags	
Rank	no of words	% of words	no of words	% of words
1	250	62.0	307	76.1
2	59	14.6	56	13.8
3	28	6.9	24	5.9

Table 6.8 : Precision of Type Identifier.

Effect of the size of manually tagged corpus

It is known that the Natural Language corpus suffers from sparseness. There are also various studies to predict the 'ideal' size of the corpus for a specific task. We state here our experience with the manually tagged compound for type identification.

With 150K corpus, the results were as follows :-

(i) Around 10% of the compounds were repeated.

(ii) The 12,796 tokens of compounds contain 2,630 types of left word and 7,106 types of right word.

(iii) The frequency distribution of highly frequent tags is shown in Table 6.10. To study the effect of fine-grained-ness we also merged the sub-types. Table 6.12 gives the frequency distribution of major tags, after merging the sub-types.

Tag	% of words
T6	28.35
Bs6	12.45
K1	9.63
Tn	8.56
Di	7.23
U	5.73

Table 6.10 : Distribution of Fine-grain-Tags

Tag	% of words
T	52.43
B	18.96
K	12.04
D	8.84
U	5.73

Table 6.12 : Distribution of Coarse-Tags

(iv) Finally the performance of type identifier on this two data was

Thus, if we consider only the 1st rank, the precision with 8 tags is 72.7% and with 55 tags, it is 63.0%. The precision increases substantially to 95.4% and 81.1% respectively if we take 1st three

	with 55 tags		with 8 tags	
Rank	no of words	% of words	no of words	% of words
1	252	63.0	291	72.7
2	44	10.9	53	13.2
3	29	7.2	38	9.5

Table 6.14 : Precision of Type Identifier.

ranks into account.

What can be concluded is, to improve the performance of type identifier, one may need to explore other parameters, other than the left or right word of a compound. For example, it might be useful to check whether the component is in neuter gender or not, whether it is a non-finite (कृदन्त) form ending in क्त or not, and so on. Such features may further improve the performance.

Chapter 7

Sanskrit compound paraphrase generator

7.1 Introduction

Paraphrase or विग्रहवाक्यम् [1] is an expression providing the meaning of a compound. In Sanskrit tradition, two types of paraphrases are discussed : 1) स्वपदविग्रहः and 2) अस्वपदविग्रहः ।

The paraphrase involving only the components of a compound, is known as स्वपदविग्रहः[2]. For instance for the compound नगरगतः (The man who has gone to the city) , the paraphrase would be नगरं गतः and this meaning is expressed by the components which are already available in the compound नगरगतः. When the paraphrase is not expressed by the components of a

[1]वृत्त्यर्थावबोधकं वाक्यं विग्रहः ।

[2]समासस्यार्थः समासघटकैः पदैर्यदि वर्ण्यते तदा स्वपदविग्रहः । तथा च - समासघटकपदसहितं वाक्यं स्वपदविग्रहः इत्यर्थः । (स्व(=समास)पदैः विग्रहः ।) -- समासः

compound, it is known as अस्वपदविग्रहः[3]. For instance the paraphrase of a compound उपकृष्णम् (Near to Krishna) is कृष्णस्य समीपम्. Here, the word समीपम् is not a component of a compound उपकृष्णम्.

Three different ways have been observed of expressing the paraphrase of a compound in Sanskrit. For instance for तपस्स्वाध्यायनिरतम् [4], the paraphrase can be expressed as :-

a. तपः च स्वाध्यायः च = तपस्स्वाध्यायौ,
तपसि च स्वाध्याये च निरतः = तपस्स्वाध्यायनिरतः,
तं तपस्स्वाध्यायनिरतम् or as

b. तपः च स्वाध्यायः च = तपस्स्वाध्यायौ,
तपस्स्वाध्याययोः निरतः = तपस्स्वाध्यायनिरतः,
तं तपस्स्वाध्यायनिरतम् or as

c. तपः च स्वाध्यायः च = तपस्स्वाध्यायौ,
तयोः निरतः = तपस्स्वाध्यायनिरतः,
तं तपस्स्वाध्यायनिरतम्

Since the second type of paraphrase is more popular and found in most of the Literature, we decided to generate the paraphrase following the second type.

7.2 Paraphrase generator

A paraphrase generator takes a well formulated tagged compound as an input and produces its paraphrase as an output. A semantically analysed compound has the following syntax.

[3]समासस्यार्थः समासघटकपदानि विहाय पदान्तरैर्यदि वर्ण्यते तदा अस्वपदविग्रहः । तथा च - समासघटकपदरहितं वाक्यम् अस्वपदविग्रहः इत्यर्थः । (अत्र एकं पदं समासघटकमेव प्रायो भवति ।) -- समासः

[4]सङ्क्षेप-रामायणम्,Bālakāṇda, sloka-1

compound : '<' component '-' component '>' tag

component : word | compound

tag : A[1-7]

| Bs[2-7] | Bs[dgpsu] | Bsm[gn] | Bv[sSU] | B[bv]

| D[is] | K[1-5] | Km | T[1-7] | T[bgmnpk]

| Tds | [ESd] | U[1-5,7]

word : [a-zA-Z]+

Sanskrit compounds tagging syntax

Note here that the compound is binary except in the case of dvandva and bahupada bahuvrihi.

We define a compound formed with the leaf nodes of a binary tree as a simple compound. Compounds with at least one component as a compound (i.e. a non-leaf node) are termed as nested compounds.

7.2.1 Paraphrase Generation

In Sanskrit compounds, as mentioned earlier, only the last component contains a case suffix[5]. Hence the major task of the paraphrase generator is to decide the gender, number and case suffix of the last component of a compound. The paraphrase of a compound varies with the type of a compound. Appendix-A gives rules for generating paraphrases for different compound types. Here are some examples :-

Example-1 Input : <दशरथ-पुत्रः>T6

Output : दशरथस्य पुत्रः = दशरथपुत्रः

The general rule for generating the paraphrase of a T6 type

[5]with an exception of an aluk samāsaḥ

compound is

<W_1 - W_2>T6 => W_1'{6} - W_2'{1} = (W_1' + W_2'){1}

Where W_1' and W_2' are the प्रातिपदिकs (nominal stem) of the words W_1 and W_2 respectively, '{6}' and '{1}' indicate the vibhakti. W_1' + W_2' on the RHS is sandhied form of the two प्रातिपदिकs W_1' and W_2'.

Example-2

Input : <पीत-अम्बरः>Bs6

Output : पीतम् अम्बरम् यस्य सः = पीताम्बरः

The general rule for generating the paraphrase is

<W_1 - W_2>Bs6 => W_1'{g}{1}, W_2'{g}{1} yat {g'}{6} tat{g'}{1}

= (W_1' + W_2'){g'}{1}

Where g' is the gender of W_2 and g is the default gender of W_2. The word अम्बरः is used in masculine here and hence its gender is masculine, which is g' while the default gender of अम्बर is neuter which is denoted by 'g' above.

7.2.2 Paraphrase generation of simple compounds

The paraphrase generation involves two major steps. In the first step we analyse the components and in the second step we generate the required word forms and construct the paraphrase. So we now describe the algorithm.

Step-A

Input : <W_1 - W_2>T_n

a. Analyse W_1, Here W_1 is an 'iic' 'in init composite' i.e. a समास-पूर्वपद. So we are interseted only in that analysis of W_1 where it can occur as a समासपूर्वपद. In case there are more than one analysis possible then we get the one which is more probable. For example, for दशरथ - two analysis are possible

i) दशरथ, पुं, iic

ii) दशरथ, नपुं, iic

Here we choose दशरथ,पुं as it is more probable. Similarly in case of राज- two analysis are possible

i) राज,पुं,iic

ii) राजन्,पुं,iic

Among these two राजन् is more probable.

To choose the more probable answer, we use a database of iics - समासपूर्वपदs. This has following 3 fields पूर्वपद form, प्रातिपदिकम् and लिङ्गम्.

Eg. राज-,राजन्,पुं

दशरथ-, दशरथ,पुं

This database is generated by extracting the समस्तपद entries form Apte's dictionary.

Let the default प्रातिपदिकम् and gender for W_1 be W_1' and g_1 respectively.

b. Analyse W_2

The analysis of W_2 pose certain problems. Just as in the case of W_1, here also multiple analysis are possible. However only those ambiguities matter us where the analyser shows analysis

with more than one gender or vibhaktis while in the context corresponding to a certain reading only one gender or vibhakti is possible. Since our paraphrase generator does not look at the context, it produces the most probable answer and has a provision to produce other answer on demand.

Another problem with the analysis of 'ife' 'in fini compositi' समास-उत्तरपद is when the समस्तपद undergoes some operations related to क्लीबत्वं then the morphological analyser should handle it. Eg. consider सप्तगङ्गम्. Here the ife is गङ्गम् which appears in neuter gender while the default gender of the word is feminine with प्रातिपदिकम् - गङ्गा.

Similar problem of change in the gender one encounters in the analysis of ife is with the Bahuvrihi compounds where the उत्तरपद assumes the gender of the object it refers to. Eg. the defualt gender of अम्बर is neuter. But in a Bahuvrihi compound, say eg. पीताम्बरः, अम्बर is in masculine since the word पीताम्बरः refers to विष्णु. So to handle these cases, we need

i) a morphological analyser which analyses words even if they are declined in some other gender. Thus this morphological analyser should be able to analyse गङ्गम् and अम्बरः, in addition to regular forms गङ्गा and अम्बरम्.

ii) a database that gives the defualt gender of a प्रातिपदिकम्.

Step-B

In this step, a paraphrase is generated. The Paraphrase has two parts :The phrase explaining the meaning and a compound word

denoting this meaning. For instance for the compound पीताम्बरः (a person who is in yellow dress) the paraphrase is पीतम् अम्बरम् यस्य सः=पीताम्बरः. In this, the पीतम् अम्बरं यस्य सः is LHS and पीताम्बरः is the RHS. The अम्बरं on the LHS has its default gender while अम्बरः on the right hand side is the gender of the object the word refers to.

Step-C Finally, if the compound word is not in the प्रथमा-विभक्ति, then appropriate pronominal phrase is also generated as

tat{g}{vibh}{num}

W_2{g}{vibh}{num}

Where g, is the gender, vibh and num are the vibhakti and number of the ifc - समास-उत्तरपद.

7.2.3 Paraphrase generation of nested compounds

In case of nested compounds, one or both the components are compounds. So we apply the above procedure repeatedly starting with the innermost compound which is a simple compound and go on simplifying it till all the compounds are covered.

7.2.4 Problem cases and thier solutions

- अलुक्समासः:-

  Only the last component of a compound has a case suffix. However as noted earlier there are exceptions typically with certain compounds whose first and intermediate components also have Vibhaktis. Such compounds are called aluk samāsa. The tagset of the Sanskrit Consortium does not mark aluk samāsas. Since the aluk samāsas are few in number, we treat as exceptional cases and produce their

output just by table lookup.

- मध्यमपद्लोपिसमासः :-

  This is a special type of compound in which some of the words in the paraphrase do not occur in its compound form. e.g. Devabrāhmaṇaḥ is a compound whose paraphrase is Devapujakaḥ brāhmaṇaḥ(a brāhmin who worships god). So to get the paraphrase of such compounds mere components are not sufficient. One should also know the context to supply the missing words. We again list out these compounds as exceptions. It is neccessary to study these compounds seperately and see if it is possible to provide some semantic criterion to provide the missing elements.

- **Special cases from** Gaṇapāṭa etc. :-

  Compounds with special paraphrases have been listed by Pāṇini separately in a list. Examples of such compounds are Mayuravaymsakaḥ, Kambojamuṇḍaḥ, Yavanamuṇḍaḥ etc. Each one of them have a special paraphrase. Readymade paraphrases of such compounds are provided.

- उपपद्समासः :-

  An upapada tatpuruṣa samāsaḥ has a verbal noun (kṛdanta) as a post component(e.g. jnaḥ and kāraḥ in Tattvajnaḥ and Kumbhakāraḥ respectively). These forms are special and occur only as bound forms in a compound. Hence, a special morphological analyser to handle these forms is built.

- **The requirement of a special morph** :-

  From generation point of view, determining the gender of

constituents is the most difficult one. For instance in उपगङ्गम् (Near to Ganges river), the second constituent of the compound viz गङ्गम् is in neuter gender, derived from the word गङ्गा (the Ganges river). The word गङ्गा is in feminine gender. Another instance from बहुव्रीहि compound is पाचिकाभार्यः (The person whose wife is a cook.). Here, the word भार्यः is the second constituent and it is in mesculine gender derived from the word भार्या (the wife). The word भार्या is in feminine gender. So the word गङ्गा and the word भार्या can be easily analysed by the morph. But when these words occur in compounds then the word गङ्गा becomes गङ्गम् and the word भार्या becomes भार्यः due to compound formation and this is place where our morphological analyser fails to analyse the words. Here, we require a special morphological analyser which can analyse these kind of words and can provide the correct stems, genders and the information of first and second components.

- **The requirement of Sandhi module** :-

  As Sandhi is mandatory in compounds, it becomes neccessary to have a Sandhi module which can join the constituents according to the Paninian theory. The Sandhi module comes in picture at the stage of paraphrase generation in RHS (Right hand side) of the paraphrase. For instance सुमित्रानन्दः (the happiness of Sumitra) and it is tagged as <सुमित्रा-आनन्दः>T6

  Here, x = सुमित्रा, y = आनन्दः and T_n = T6 (षष्ठी-तत्पुरुषः)

  सुमित्रायाः आनन्दः = सुमित्रानन्दः is the complete paraphrase of <सुमित्रा-आनन्दः>T6.

In paraphrase

LHS = सुमित्रायाः आनन्दः

and

RHS = सुमित्रानन्दः

Now, for generating LHS a module is not required but when it goes to RHS, it needs a Sandhi module to generate सुमित्रानन्दः.

- **Special treatment of Dvandva compounds**

The gender and number of द्वन्द्व compounds depend on number of components and sometimes even on the semantics. For instance बककाकौ(The Crane (bird) and the Crow) and बककाकाः (The Crane or the group of Cranes and the group of Crows). The instance बककाकौ contains only two components बक and काक and it is in dual number. So simply by looking at the number of components and the number of compound, the paraphrase can be बकः च काकः च = बककाकौ which gives the meaning that only two single birds are there. But बककाकाः is not the same case, although the compound contains only two components and has the same gender, the word is now in plural not in dual. It is becuase of the involved semantics. Here either of both the words बक and काक denote a जाति and not an individual. Hence the paraphrase in this case is बकाः च काकाः च = बककाकाः or बकः च काकाः च = बककाकाः. Now the question is how to treat such compounds ? Where is the information that the component refers to a जाति and not a व्यक्ति. One may argue reterning to the सूत्र -"जातिरप्राणिनाम्" (2.4.6) that one may list such words and handle separately. But as we see above both <बक-काकाः> as well as <बक-काकौ> is possible. In one case it refers to the जाति and in another

case व्यक्ति. So one may then argue that the information is in the form itself. Yes it is true that the information is in the word form. But when there are 3 or more than 3 components or if such components are part of another compound then this information will not be available. In such cases we asume that the word refers to a व्यक्ति.

7.3 Evaluation

The evaluation here is simple one. It does not involve any precision or recall figures, but just the % cases that produce correct paraphrase. We tested 200 simple compounds and 100 nested compounds. 89% simple compounds and 80% nested compounds paraphrases were correct.

Chapter 8

Conclusion

The present work is an effort towards building an Automatic Computational Sanskrit Compound Processing. We observed four major tasks in compound analysis: 1) Segmentation, 2) Constituency parsing, 3) Type-identification and 4) Paraphrasing. Except the last task each of the first three involves non-determinism and hence we used statistical data for prioritizing the solutions. We have used sandhi rules in reverse to split and then we have used probability for prioritizing the solutions. For Constituency Parsing, we have used only heuristic rules. For Type-identification of compound, the Paninian rules are applied and wherever the rules fail, we have used the statistical data.

Limitataions of Sanskrit Compound Processor

- The input for the system should be a complete compound word. It should not be a sentence or the word which contains only sandhi.
- The automatic segmentation can be done only for those the compound words which contain four or less than four components.

- The automatic constituency parser does not handle the द्वन्द्व compounds at this level.
- An automatic type-identification of compound can be done only for those compounds which contain only two components.
- For paragraphing of मध्यमपद-लोपि and उपपद-तत्पुरुष compound are not possible at this level.

Keeping these limitataions in mind one can use the system.

Future directions

The segmentation of components can be improved further by using the sandhi rules which are used only in compounds. Constituency parsing can be done more effectively if rules of Panini are also applied.

The present work can be a model for other modern Indian languages. The processing of compounds is still untouched in modern Indian languages. The process which we have described in our work for Sanskrit can be used for other modern Indian languages also.

Appendices

A - Table of paraphrase rules

अव्ययीभावः	
1	<x-y>A1 => y6 fx where f maps x to the noun with same semantic content. A function f needs to be defined.
2	<x-y>A2 => x3 विपरीतं वृत्तम्
3	<x-y>A4 => x6 y6 समाहारः
4	<x-y>A5 => x'1 y1 यस्मिन् देशे x' y' इत्यनयोः समानलिङ्गं स्यात् इति।
5	<x-y>A6 => x6 y6 समाहारः यदि x=द्वि स्यात् तर्हि x एवं y इत्यनयोः प्रयोगः द्विवचने स्यात् ।
6	<x-y>A7 => x6 y
तत्पुरुषः	
7	<x-y>T_n => xn y $2 \leq n \leq 7$
8	<x-y>Tn => xn y
9	<x-y>Tds => x6 ;ba y6 ;ba समाहारः
10	<x-y-z>Tb => x1 y1 z1
कर्मधारयः	

11	<x-y>K1 => x1 तत् y1 च
12	<x-y>K2 => x1 चy1 च
13	<x-y>K3 => x1 च असौ y1 च
14	<x-y>K4 => x1 इव y1
15	<x-y>K5 => x1 y1 इव
16	<x-y>K6 => x1 एव y1
17	<x-y>K7 => x1 इति y1
बहुव्रीहिः	
18	<x-y>Bsd => x6 च y6 च यदन्तरालम्
19	<x-y>Bsp => x3 च y3 च प्रहृत्य इदं युद्धं प्रवृत्तम्
20	<x-y>Bsg => x7 y7 गृहीत्वा इदं युद्धं प्रवृत्तम्
21	<x-y>Bsmn => x' y1 यस्य
22	<x-y>Bss => x1 वा y1 यस्य
23	<x-y>Bsu => x1 इव y1 यस्याः
24	<x-y>Bv => x y1 यस्य
25	<x-y>Bvs => y6 x' ये सन्ति ते
26	<x-y>BvS => y3 सह
27	<x-y>BvU => x6 इव y यस्य
द्वन्द्वः	
28	<x-y>Di => x1 च y1 च
29	<x-y>Ds => x1 च y1 च
30	<x-y>S => y1 x1
31	<x-y>d => x y

B - Table of Semantic classifications

Semantic classifications of Sanskrit compounds					
अव्ययीभाव			तत्पुरुष		
1	अव्यय-पूर्वपद-अव्ययीभाव	A1	1	प्रथमातत्पुरुष	T1
2	अव्यय-उत्तरपद-अव्ययीभाव	A2	2	द्वितीयातत्पुरुष	T2
3	तिष्ठद्गुप्रभृति-अव्ययीभाव	A3	3	तृतीयातत्पुरुष	T3
4	संख्यापूर्वपद-नद्युत्तरपद-अव्ययीभाव	A4	4	चतुर्थीतत्पुरुष	T4
5	नद्युत्तरपद-अन्यपदार्थसंज्ञायाम्	A5	5	पञ्चमीतत्पुरुष	T5
6	संख्यापूर्वपद-वंश्योत्तरपद-अव्ययीभाव	A6	6	षष्ठीतत्पुरुष	T6
7	पारे-मध्ये-पूर्वपदषष्ठ्युत्तरपद-अव्ययीभाव	A7	7	सप्तमीतत्पुरुष	T7
बहुव्रीहि			8	नञ्-तत्पुरुष	Tn
1	द्वितीयार्थबहुव्रीहि	Bs2	9	प्रादि-तत्पुरुष	Tp
2	तृतीयार्थबहुव्रीहि	Bs3	10	कु-तत्पुरुष	Tk
3	चतुर्थ्यर्थबहुव्रीहि	Bs4	11	गति-तत्पुरुष	Tg
4	पञ्चम्यर्थबहुव्रीहि	Bs5	12	तद्धितार्थद्विगु	Tdt
5	षष्ठ्यर्थबहुव्रीहि	Bs6	13	उत्तरपदद्विगु	Tdu
6	सप्तम्यर्थबहुव्रीहि	Bs7	14	समाहारद्विगु	Tds
7	दिग्वाचक-बहुव्रीहि	Bsd	15	उपपद	U

8	संख्योभयपद-बहुव्रीहि	Bss	16	द्वितीयोपपद-तत्पुरुष	U2
9	उपमानपूर्वपद-बहुव्रीहि	Bsu	17	तृतीयोपपद-तत्पुरुष	U3
10	प्रहरणविषयक-बहुव्रीहि	Bsp	18	चतुर्थ्योपपद-तत्पुरुष	U4
11	ग्रहणविषयक-बहुव्रीहि	Bsg	19	पञ्चम्योपपद-तत्पुरुष	U5
12	सङ्ख्योत्तरपद-व्यधिकरण-बहुव्रीहि	Bvs	20	सप्तम्योपपद-तत्पुरुष	U7
13	सहपूर्वपद-व्यधिकरण-बहुव्रीहि	BvS	21	मयूरव्यंस्कादि	Tm
14	प्रादि-व्यधिकरण-बहुव्रीहि	Bvp	22	बहुपद-तत्पुरुष	Tb
15	उपमानपूर्वपद-व्यधिकरण-बहुव्रीहि	BvU	कर्मधारय		
16	नञ्-बहुव्रीहि	Bsmn	1	विशेषण-पूर्वपद-कर्मधारय	K1
17	बहुपद-बहुव्रीहि	Bb	2	विशेषण-उत्तरपद-कर्मधारय	K2
द्वन्द्व			3	विशेषण-उभयपद-कर्मधारय	K3
1	इतरेतर-द्वन्द्व	Di	4	उपमान-पूर्वपद-कर्मधारय	K4
2	समाहार-द्वन्द्व	Ds	5	उपमान-उत्तरपद-कर्मधारय	K5
3	एकशेष-द्वन्द्व	E	6	अवधारणापूर्वपद-कर्मधारय	K6
अन्य (others)			7	सम्भावनापूर्वपद-कर्मधारय	K7
1	द्विरुक्ति	d	8	मध्यमपदलोपिकर्मधारय	Km
2	केवल-समास	S	-		

C - List of aphorisms

List of implemented aphorisms

1. "अव्ययं विभक्ति-समीप-समृद्धि-व्यृद्ध्यर्थाभावात्यया-सम्प्रति-शब्दप्रादुर्भाव..." (2-1-6)
2. "यावदवधारणे" (2-1-8)
3. "सुप्प्रतिना मात्रार्थे" (2-1-9)
4. "अक्षशलाकासंख्याः परिणा" (2-1-10)
5. "अपपरिबहिरञ्चवः पञ्चम्या" (2-1-12)
6. "आङ् मर्यादाऽभिविध्योः" (2-1-13)
7. "लक्षणेनाऽभिप्रती आभिमुख्ये" (2-1-14)
8. "अनुर्यत्समया" (2-1-15)
9. "यस्य चायामः" (2-1-16)
10. "तिष्ठद्गुप्रभृतीनि च" (2-1-17)
11. "पारे मध्ये षष्ठ्या वा" (2-1-18)
12. "संख्या वंश्येन" (2-1-19)
13. "नदीभिश्च" (2-1-20)
14. "द्वितीया श्रितातीतपतितगतात्यस्तप्राप्तापन्नैः" (2-1-24)
15. "स्वयं क्तेन" (2-1-25)
16. "खट्वा क्षेपे" (2-1-26)

17. "सामि" (2-1-27)
18. "कालाः" (2-1-28)
19. "पूर्वसदृशसमोनार्थकलहनिपुणमिश्रश्लक्ष्णैः" (2-1-31)
20. "चतुर्थी तदर्थार्थबलिहितसुखरक्षितैः" (2-1-36)
21. "पञ्चमी भयेन" (2-1-37)
22. "अपेतापोढमुक्तपतितापत्रस्तैरल्पशः" (2-1-38)
23. "स्तोकान्तिकदूरार्थकृच्छ्राणि क्तेन" (2-1-39)
24. "सप्तमी शौण्डैः" (2-1-40)
25. "सिद्धशुष्कपक्वबन्धैश्च" (2-1-41)
26. "ध्वाङ्क्षेण क्षेपे" (2-1-42)
27. "क्तेनाहोरात्रावयवाः" (2-1-45)
28. "तत्र" (2-1-46)
29. "पात्रेसमितादयश्च" (2-1-48)
30. "पूर्वकालैकसर्वजरत्पुराणनवकेवलाः समानाधिकरणेन" (2-1-49)
31. "पापाणके कुत्सितैः" (2-1-54)
32. "उपमितं व्याघ्रादिभिः सामान्याप्रयोगे" (2-1-56)
33. "विशेषणं विशेष्येण बहुलम्" (2-1-57)
34. "पूर्वापरप्रथमचरमजघन्यसमानमध्यमध्यमवीराश्च" (2-1-58)
35. "श्रेण्यादयः कृतादिभिः" (2-1-59)
36. "क्तेन नञ्विशिष्टेनानञ्" (2-1-60)
37. "सन्महत्परमोत्तमोत्कृष्टाः पूज्यमानैः" (2-1-61)
38. "वृन्दारकनागकुञ्जरैः पूज्यमानम्" (2-1-62)
39. "कतरकतमौ जातिपरिप्रश्ने" (2-1-63)
40. "किं क्षेपे" (2-1-64)
41. "पोटायुवतिस्तोककतिपयगृष्टिधेनुवशावेहद्वष्कयणीप्रवक्तृश्रोत्रिय..." (2-1-65)

42. "प्रशंसावचनैश्च" (2-1-66)
43. "युवा खलतिपलितवलिनजरतीभिः" (2-1-67)
44. "वर्णो वर्णेन" (2-1-69)
45. "चतुष्पादो गर्भिण्या" (2-1-71)
46. "मयूरव्यंसकादयश्च" (2-1-72)
47. "अर्धं नपुंसकम्" (2-2-2)
48. "द्वितीयतृतीयचतुर्थतुर्याण्यन्यतरस्याम्" (2-2-3)
49. "नञ्" (2-2-6)
50. "ईषदकृता" (2-2-7)
51. "याजकादिभिश्च" (2-2-9)
52. "कुगतिप्रादयः" (2-2-18)
53. "संख्ययाव्ययासन्नदूराधिकसंख्याः संख्येये" (2-2-25)
54. "दिङ्नामान्यन्तराले" (2-2-26)
55. "तेन सहेति तुल्ययोगे" (2-2-28)

List of non-implemented aphorisms

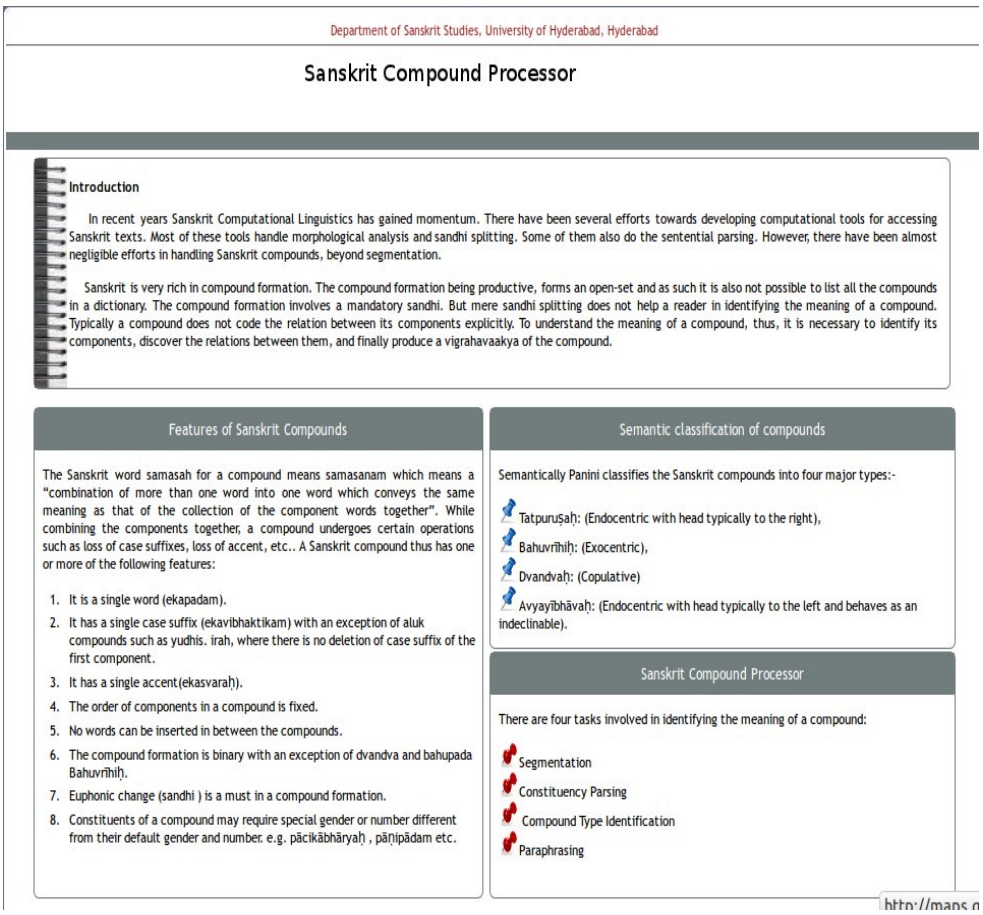
1. "यथाऽसादृश्ये" (2-1-7)
2. "अत्यन्तसंयोगे च" (2-1-29)
3. "तृतीया तत्कृतार्थेन गुणवचनेन" (2-1-30)
4. "कर्तृकरणे कृता बहुलम्" (2-1-32)
5. "कृत्यैरधिकार्थवचने" (2-1-33)
6. "अन्नेन व्यञ्जनम्" (2-1-34)
7. "भक्ष्येण मिश्रीकरणम्" (2-1-35)
8. "कृत्यैरृणे" (2-1-43)
9. "संज्ञायाम्" (2-1-44)

10. "क्षेपे" (2-1-47)
11. "दिक्संख्ये संज्ञायाम्" (2-1-50)
12. "तद्धितार्थोत्तरपदसमाहारे च" (2-1-51)
13. "संख्यापूर्वो द्विगुः" (2-1-52)
14. "कुत्सितानि कुत्सनैः" (2-1-53)
15. "उपमानानि सामान्यवचनैः" (2-1-55)
16. "कृत्यतुल्याख्या अजात्या" (2-1-68)
17. "पूर्वापराधरोत्तरमेकदेशिनैकाधिकरणे" (2-2-1)
18. "नित्यं क्रीडाजीविकयोः" (2-2-17)
19. "उपपदमतिङ्" (2-2-19)
20. "अनेकमन्यपदार्थे" (2-2-24)
21. "चार्थे द्वन्द्वः" (2-2-29)

List of Exception aphorisms

1. "अन्यपदार्थे च संज्ञायाम्" (2-1-21)
2. "कुमारः श्रमणादिभिः" (2-1-70)
3. "प्राप्तापन्ने च द्वितीयया" (2-2-4)
4. "कालाः परिमाणिना" (2-2-5)
5. "अमैवाव्ययेन" (2-2-20)
6. "तृतीयाप्रभृतीन्यन्यतरस्याम्" (2-2-21)
7. "तत्र तेनेदमिति सरूपे" (2-2-27)

D - Screenshots



Department of Sanskrit Studies, University of Hyderabad, Hyderabad

Sanskrit Compound Processor

Introduction

In recent years Sanskrit Computational Linguistics has gained momentum. There have been several efforts towards developing computational tools for accessing Sanskrit texts. Most of these tools handle morphological analysis and sandhi splitting. Some of them also do the sentential parsing. However, there have been almost negligible efforts in handling Sanskrit compounds, beyond segmentation.

Sanskrit is very rich in compound formation. The compound formation being productive, 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, thus, it is necessary to identify its components, discover the relations between them, and finally produce a vigrahavaakya of the compound.

Features of Sanskrit Compounds

The Sanskrit word samasah for a compound means samasanam which means a "combination of more than one word into one word which conveys the same meaning as that of the collection of the component words together". While combining the components together, a compound undergoes certain operations such as loss of case suffixes, loss of accent, etc.. A Sanskrit compound thus has one or more of the following features:

1. It is a single word (ekapadam).
2. It has a single case suffix (ekavibhaktikam) with an exception of aluk compounds such as yudhis. irah, where there is no deletion of case suffix of the first component.
3. It has a single accent(ekasvaraḥ).
4. The order of components in a compound is fixed.
5. No words can be inserted in between the compounds.
6. The compound formation is binary with an exception of dvandva and bahupada Bahuvrīhiḥ.
7. Euphonic change (sandhi) is a must in a compound formation.
8. Constituents of a compound may require special gender or number different from their default gender and number. e.g. pācikābhāryaḥ , pāṇipādam etc.

Semantic classification of compounds

Semantically Panini classifies the Sanskrit compounds into four major types:-

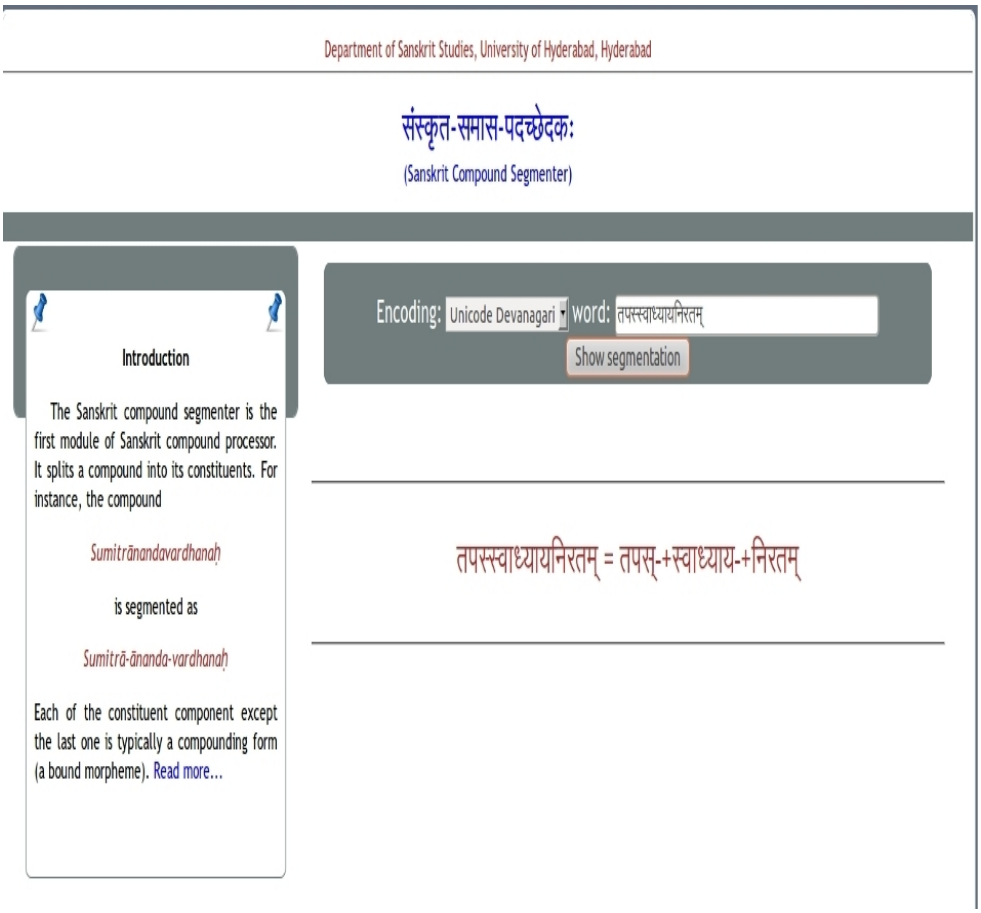
- Tatpuruṣaḥ: (Endocentric with head typically to the right),
- Bahuvrīhiḥ: (Exocentric),
- Dvandvaḥ: (Copulative)
- Avyayībhāvaḥ: (Endocentric with head typically to the left and behaves as an indeclinable).

Sanskrit Compound Processor

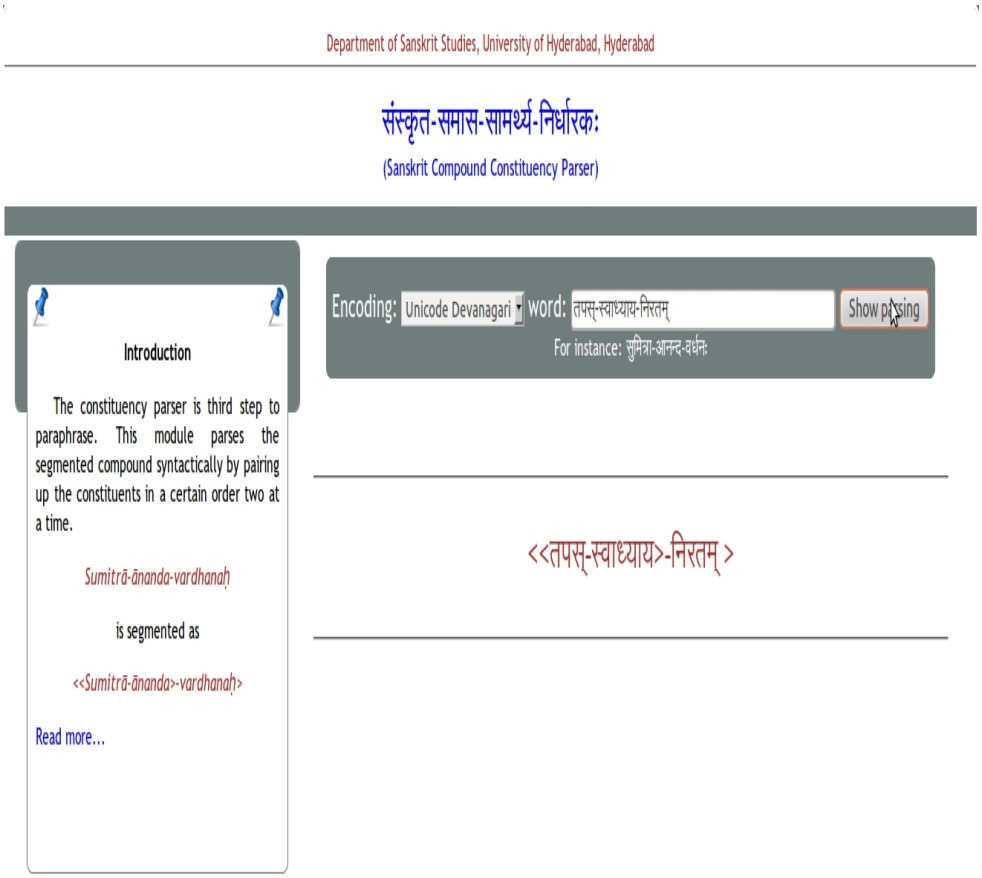
There are four tasks involved in identifying the meaning of a compound:

- Segmentation
- Constituency Parsing
- Compound Type Identification
- Paraphrasing

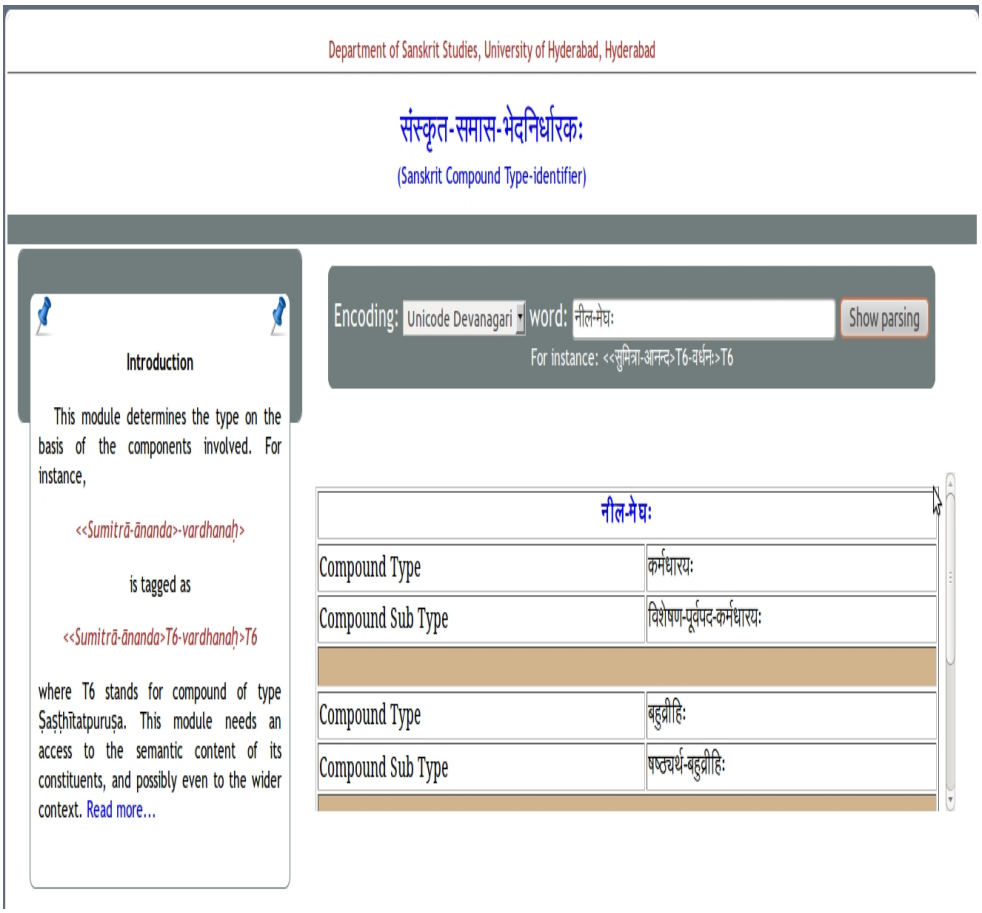
Home page of Sanskrit Compound Processor



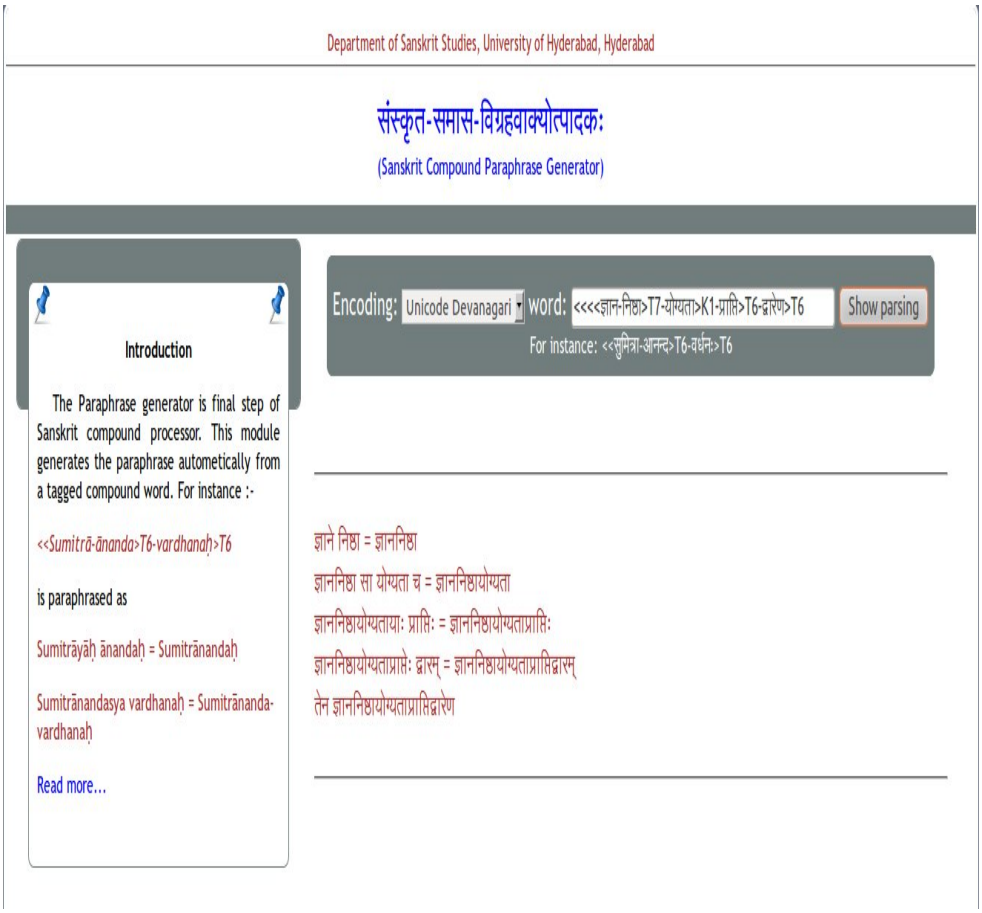
Screenshot of Segmenter



Screenshot of Constituency Parser



Screenshot of Type-Identifier



Screenshot of Paraphrase Generator

E-Bibliography

Articles/Papers/Books

1. Bharati, A., Kulkarni, A.P., and Sheeba, V. 2006 . *Building a wide coverage Sanskrit Morphological Analyser : A practical approach*. In : The First National Symposium on Modelling and Shallow Parsing of Indian Languages, IIT-Bombay.
2. Bhat, G.M. 2006 : Samāsaḥ. Samskrita Bharati, Bangalore, Karnataka.
3. Fortes, F.C.L., and Roxas, R.E.O. : *Optimality Theory in Morphological Analysis* In : National Natural Language Processing Research Symposium, January 2004.
4. Fosler, J.E. 1996 : *On Reversing the Generation Process in Optimality Theory*. In : Proceedings of the Association for Computational Linguistics.
5. Gillon, B.S. 2007 :*Exocentric Compounds in Classical Sanskrit*. In : Proceeding of the First International Symposium on Sanskrit Computational Linguistics(SCLS-2007), Paris, France.
6. Gillon, B.S. 2009 : *Tagging Classical Sanskrit Compounds*. In : Sanskrit Computational Linguistics 3, pages 98-105, Springer-Verlag

LNAI 5406.

7. Hellwig, O. 2007 : *Sanskrit Tagger : A Stochastic Lexical and POS Tagger for Sanskrit*. In : Sanskrit Computational Linguistics 1 & 2, pages 266-277, Springer-Verlag LNAI 5402.
8. Hellwig, O. 2009 : *Extracting Dependency Trees from Sanskrit Texts*. In : Sanskrit Computational Linguistics 3, pages 106-115, Springer-Verlag LNAI 5406.
9. Huddleston, R. and Pullum, G.K. 2006 : Cambridge Grammar English Language (Chapter-19 only). Cambridge University Press, United Kingdom.
10. Huet, G. 2006 : *Lexicon-directed Segmentation and Tagging of Sanskrit*. XIIth World Sanskrit Conference, Helsinki, Finland, Aug. 2003. In Themes and Tasks in Old and Middle Indo-Aryan Linguistics, Eds. Bertil Tikkanen and Heinrich Hettrich. Motilal Banarsidass, Delhi, pp. 307-325.
11. Huet, G. 2006 : *Shallow syntax analysis in Sanskrit guided by semantic nets constraints*. In : Proceedings of International Workshop on Research Issues in Digital Libraries, Kolkata.
12. Huet, G. 2007 : *Formal structure of Sanskrit text : Requirements analysis for a Mechanical Sanskrit Processor*. In : Sanskrit Computational Linguistics 1 & 2, pages 162-199, Springer-Verlag LNAI 5402.
13. Huet, G. 2009 : *Sanskrit Segmentation*. In : South Asian Languages Analysis Roundtable XXVIII, Denton, Ohio.
14. Jha, B.G. 1990 : Samāsa-sandarśikā. Chowkhamba Surabharati

Prakashan, Varanasi, UP.

15. Jha, G.N. and Mishra, S.K. 2007 : *Semantic Processing in Panini's Karaka System*. In : Sanskrit Computational Linguistics 1 & 2, pages 239-252, Springer-Verlag LNAI 5402.
16. Joshi, S.D. and J.A.F. Roodbergen. 1969 : The Vaiyākaraṇa-mahābhāṣya (avyayībhāvatatpuruṣāhnika). University of Poona, Poona, Maharastra.
17. Joshi, S.D. and J.A.F. Roodbergen. 1973 : The Vaiyākaraṇa-mahābhāṣya (Tatpuruṣāhnika). University of Poona, Poona, Maharastra.
18. Joshi, S.D., J.A.F. Roodbergen. 1996 : The Aṣṭādhyāyī of Pāṇini - Volume V and VI. Sahitya Academy, New-Delhi (India).
19. 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.
20. Kulkarni, A.P., Shukla, D. 2009 : *Sanskrit Morphological Analyser : Some Issues*. In : To appear in Bh.K Festschrift volume by LSI.
21. Kumar, A., Mittal, V., Kulkarni, A.P. 2010 : *Sanskrit Compound Processor*. In : Proceedings of 4i-SCLS 2010 : 4th International Sanskrit Computational Linguistics Symposium, Springer-Verlag LNAI 6465.
22. Mahavira. June 1978 : Pāṇini as Grammarian (With special reference to compound formation). Bharatiya Vidya Prakashan [Delhi - Varanasi], India.
23. Mimamsaka, Yudhisthir : Mahābhāṣyam (with Hindi commentory) -

I, II and III parts. Ramlal Kapur Trust, Sonepat, Haryana.

24. Mittal, V. 2010 : *Automatic Sanskrit Segmentizer using Finite State Transducers*. In : Proceeding of Association for Computational Linguistics - Student Research Workshop.
25. Murty, M.S. 1974 : Sanskrit Compounds-A Philosophical Study. Chowkhamba Sanskrit Series Office, Varanasi(India).
26. Pande, G.D. : Aṣṭādhyāyī of Pāṇini. Chowkhamba Surabharati Prakashan, Varanasi, UP.
27. Pandit Ishvarachandra. 2004 : Aṣṭādhyāyī. Chaukhamba Sanskrit Pratisthan, Delhi.
28. Ramakrishnamacharya, K.V. 2010 : Bhūṣaṇasāra-tattvaprakāśikā. Rashtriya Sanskrit Vidyapeetham, Tirupati, AP.
29. Sarma, E.R.S. 1960 : Maṇikaṇa : A Navya-Nyāya Manual. The Adyar Library and research Centre, Madras.
30. Scharf, P.M. 2009 : Levels in Pāṇini's Aṣṭādhyāyī. In : Sanskrit Computational Linguistics 3, pages 66-77, Springer-Verlag LNAI 5406.
31. Sharma, Vasudeva L.S.P. : The Siddhānta Kaumudī (With Tatvabodini commentory). Tukaram Jivaji Press, Bombay.
32. Shastri, Pt. Guru Prasad. 2006 : Vyākaraṇa-mahābhāṣyam (Only Samāsaprakaraṇam). Rashtriya Sanskrit Sansthan, New-Delhi.
33. Subrahmanyam, K. : Four Vṛittis in Pāṇini. Library of Congress Control Number : 2001445310.
34. Tarkavachaspati, Taranatha, 1812-1885 : Vacaspatyam. Chowkhamba Sanskrit Series, Varanasi, UP.

35. Tripathi, Vijay Prasad. 1991 : Samāsa-vṛtti-vimarśaḥ. Sampurnananda Sanskrit Vishvavidyalaya, Varanasi, UP.
36. Vasu, Srisha Chandra. 1891 : The Aṣṭādhyāyī of Pāṇini (Translated into English). Indian Press, Allahabad, UP.
37. Vasu, Srisha Chandra : The Siddhānta Kaumudī of Bhattoji Dikṣita. Motilal Banarsidas Publishers, New Delhi.
38. Yuret, D., Biçici, E. 2009 : *Modeling Morphologically Rich Languages Using Split Words and Unstructured Dependencies*. In : ACL-IJCNLP, Singapore.

Websites

1. Sanskrit Heritage :
   http://sanskrit.inria.fr/
2. Samsaadhanii :
   http://sanskrit.uohyd.ernet.in/scl/
3. The Computational Linguistics R&D :
   http://sanskrit.jnu.ac.in/index.jsp
4. Digital Corpus of Sanskrit :
   http://kjc-fs-cluster.kjc.uni-heidelberg.de/dcs/index.php
5. Sanskrit Deepika :
   http://www.sanskritdeepika.org/index.php ?option=com_content &task=category&sectionid=7&id=39&Itemid=42

Computer tools/softwares

1. Pāṇinīyavyākaraṇodāharaṇakoṣaḥ : La grammaire paninéenne par ses exemples. The Paninian grammar through its examples. Vol.

2. Samāsaprakaraṇam Le livre des mots composés. The book of compound words. -- . Grimal, V. Venkataraja Sarma, S. Lakshminarasimham, (Rashtriya Sanskrit Vidyapeetha Series no 150 ; Collection Indologie no 93.2, RSV, Tirupati / EFEO / IFP, 2007, xviii, 834 p.

2. Ganakastadhyayi (Windows version) :
http ://www.taralabalu.org/panini/