TAGGING CLASSICAL SANSKRIT COMPOUNDS

Brendan S. Gillon
McGill University
Montreal, Quebec

Abstract. The paper sets out a *prima facie* case for the claim that the classification of Sanskrit compounds in Pāṇinian tradition can be retrieved from a very slight augmentation of the usual enriched context free rules.

Key words: *Aṣṭādhyāyī*, Classical Sanskrit, compounds, context free rules, Pāṇini.

1 Introduction

The aim of this paper is to make a *prima facie* case that the information pertaining to the grammar of compounds in Classical Sanskrit captured in their classification by the Pāṇinian tradition can be retrieved from a very slight augmentation of the usual enriched context free rules used by generative linguists. To make the complete case would require much more space than is available here.

I shall proceed as follows. First, I shall remind the reader of the general properties of compounds in Classical Sanskrit. Second, I shall set out what I mean by enriched context free rules. Third, I shall review the classification of compounds by the Pāṇinian tradition and show, for each category, how that category can be retrieved from the structure assigned to a compound by the enriched context free rules.

2 General Properties

The following are generally acknowledged regularities which the compounds of Classical Sanskrit exhibit.

1. Compounds are subject to the inflectional and derivational morphological forms of simple words (A 2.4.71; A 6.3.1; MBh to A 2.1.1, i.e., Kielhorn (ed) [5] v. I,p. 362.5; Whitney [8] §1246; and Cardona [1] pp. 264-265). In particular, inflection occurs at the end of compounds, not within them; derivational suffixes can be added as easily to compounds as they can be to words.

2. The accentuation of a compound is that of a simple word, not that of a phrase (A 6.1.158; MBh to A 2.1.1, i.e., Kielhorn (ed) [5] v. I, pp. 362.8-9; Whitney [8] §1246).

3. Constituents of a compound, unlike constituents of phrases, have a fixed linear order (A 2.2.30; MBh to A 2.1.1, i.e., Kielhorn (ed) [5] v. I, p. 362.8; Cardona [1] pp. 261-264). In general, whereas no two immediate constituents of a compound can be transposed and the sense of the compound retained for its members; any two immediate constituents of a phrase can be transposed and the sense of the phrase retained.

4. Inflected words, which are external to a compound, are not construed with uninflected constituents subordinate within it (MBh to A 2.1.1).

5. A compound is usually analyzable into two immediate constituents (A 2.1.4); and if there is a head, it is the second immediate constituent (A 1.2.43; A 2.2.30; ; Whitney [8] §1246; Cardona [1] pp. 261-263).

6. Compounds are of unbounded complexity (Whitney [8] §1248).

7. A compound has a typical, and for Pāṇini, a canonical, phrasal paraphrase (*vigraha-vākya*) such that, if a compound has the form $[\mathrm{C\ D}]_i$ then its phrasal paraphrase has the form C_j D_i (where i and j denote one of the seven Sanskrit cases). Moreover, the head of a canonical phrasal paraphrase is the head of the compound being paraphrased.

The first four regularities make it plausible that lexical and phrasal syntax are distinct. The fifth and sixth regularities show that compounds in Classical Sanskrit are binary branching and recursive. Such are the kinds of structures which one would expect the enriched context free rules of the sort given by Selkirk (1982), among others, for English would generate.

To be sure, there are exceptions to these regularities, but happily they are not productive. For example, some compounds, such as conjunctive compounds (*dvandva* compounds) formed from the names of gods, do not have the accent of simple words. Moreover, there are cases where an inflectional affix occurs on a subordinate constituent within a compound (A 6.3.7-8), so-called *aluk* compounds, or within lexical derivation (A 6.3.17). However, these cases are not considered productive by any Sanskritist and they are best treated as items to be listed in the lexicon. (For examples and discussion, see Whitney [8] §1250 and Cardona [1], pp. 264-265.)

3 Enriched Context Free Rules

It is useful to distinguish between descriptive grammars and generative grammars. Descriptive grammars are those which state regularities which do not aim to generate all and only the well-formed expressions of the language. Traditional grammars of European languages and teaching grammars of various languages around the world are examples of such grammars. Generative grammars are those grammars which aim, on the basis of a lexicon and a set of rules, to generate all and only the acceptable expressions of a language. Pāṇini's *Aṣṭādhyāyī* is such a grammar, for it aims to do precisely that for Classical Sanskrit. Grammars of the American structuralists, what they called *constituency grammars*, are also examples of such grammars.

These grammars, though generative, were informal grammars. The first attempt to define and characterize formal grammars was undertaken by Chomsky [3][1], who distinguished regular grammars, from context free grammars, from context sensitive grammars, from unrestricted grammars. Chomsky [2] claimed that the constituency grammars of the American structuralists are properly formalized as context free grammars. However, as most formally minded linguists now recognize, this is not true. To be sure, American structuralist linguists did use rules which are instantiations of context free rules, but the rules they used had richer content, which included the use of features and of structured category labels.

Context free grammars can be viewed from a variety of perspectives. Initially, they were regarded as rewrite rules. The correlated labelled trees,

[1] This publication was based on work done before 1957.

then, represented equivalence classes of derivations. Eventually, linguists abandoned the view of context free grammars as sets of rewrite rules and adopted the view instead they characterized the order which the morphemes in a complex expression bear to one another. This is the view I shall take in what follows, though I suspect that nothing crucial depends on which of the two views one opts for. The latter view is simply the one which most linguists find more congenial.

It is now generally recognized that the labels for categories of expressions, features and subcategorization frames are all required in a grammar of a language. Enriched context free rules, then, are context free rules enriched with such additional structure. Let me state, then, some of the enrichments required for the treatment of expressions of Classical Sanskrit.[2] To begin with, one requires feature specification for case, number and gender. We shall not introduce these here, as they do not bear directly on the range of data to be addressed in this paper. In addition, one requires structured category labels. Basically, one requires labels for adjective, adverb, noun, preposition, verb and clause. However, these must be enriched so as to distinguish between stems, words (inflected stems) and phrases. The phrasal labels include A^1 (inflected adjective), A^2 (adjective phrase), N^1 (inflected noun), N^2 (noun phrase), V^1 (inflected verb), V^2 (verb phrase), P^1 (preposition), P^2 (prepositional phrase), D^1 (adverb), D^2 (adverbial phrase) and S (clause). The remaining labels, which are labels for stems, are A^0 (adjective stem), N^0 (noun stem), P^0 (preposition stem)[3] and V^0 (verb stem). (Since phrasal syntax plays only an incidental role here, I shall drop the superscripts and, unless otherwise specified X is short of X^0.) All compound stems will then have a label X (X^0). Subcategorization information is used to capture what morphologists call *bound morphemes*. I shall indicate bound morphemes with the linguist's customary use of a hyphen, preceding the morpheme, when the morpheme is suffixed to another constituent, and succeeding the morpheme, when it is prefixed. (How subcategorization is to be handled is a much more complex matter, involving many facets of the grammar not discussed here.)

[2] It is such enrichments which are at the heart of such grammars as Generalized Phrase Structure Grammar (GPSG) and Head Driven Phrase Structure Grammar (HPSG).

[3] There is probably no need to distinguish between preposition stem and prepositional word.

4 Traditional classification

Pāṇini's treatment of compounds in his *Aṣṭādhyāyī* is one familiar to contemporary linguists. Each compound is paired with a canonical phrasal paraphrase (*vigraha-vākya*). Underlying the compound and its canonical phrasal paraphrase is a string of symbols, which denote what the compound and its paraphrase denote. At a suitable point in the derivation, elements corresponding to the inflection of words in the paraphrase are optionally deleted. As a result of this correspondence between compound and canonical paraphrase, it is possible to classify the compounds using properties of the paraphrases. This, then, is the basis for the classification of compounds used in the later Pāṇinian tradition. Here is the classification:

1. *aluk*
2. *luk*
    - (a) *avyayībhāva*
    - (b) *dvandva*
    - (c) *tatpuruṣa*
        - i. *nañ tatpuruṣa*
        - ii. *prādi tatpuruṣa*
        - iii. *upapada tatpuruṣa*
        - iv. *vibhakti tatpuruṣa*
        - v. *karmadhāraya*
            - A. *viśeṣaṇa-pūrva-pada-karmadhāraya*
            - B. *viśeṣaṇa-uttara-pada-karmadhāraya*
            - C. *viśeṣaṇa-ubhaya-pada-karmadhāraya*
            - D. *upamāna-pūrva-pada-karmadhāraya*
            - E. *avadhāraṇa-pūrva-pada-karmadhāraya*
            - F. *upamāna-uttara-pada-karmadhāraya*
    - (d) *bahuvrīhi compounds*

The question is: can one recover from the analysis of a compound by enriched context free rules the classification of the compound within this schema. The answer seems to be yes. However, the full case cannot be made here, both because the data to be surveyed demands a paper much longer than what is permitted and because the problem of compound analysis is intimately connected with other aspects of the grammar whose precise treatment remains either obscure or undecided. Nonetheless, I shall sketch out the basic ideas, filling in, as required, ancillary assumptions.

The simplest cases are the *aluk* compounds. These are the compounds in which the left sister is inflected. Consider, for example, *ātmanepada*. It is analyzed as [$_N$ [$_{N^1}$ *ātmane*] [$_N$ *pada*]].[4] The fact that the analysis contains a bracket labelled with N^1 as a left sister to a bracket labelled with N is both necessary and sufficient to identify the compound as an *aluk* compound.

All other compounds are *luk* compounds, meaning the left sister constituents are all stems, either bound or unbound. Within the *luk* compounds, *avyayībhāva* compounds are easily identified by their analysis. They are compound stems inflected as adverbs, as is, for example, [$_{D^1}$ [$_P$ *upari*] [$_N$ *bhūmi*]].

Also easily identifiable from their parses are *nañ-tatpuruṣa* compounds, *upapada-tatpuruṣa* compounds and *prādi-tatpuruṣa* compounds. *Nañ-tatpuruṣa* compounds are compounds prefixed with the bound morpheme *a-* or *an-*. Thus, for example, it is evident that [$_N$ [$_A$ *a-*] [$_N$ *brāhmaṇa*]] and [$_N$ [$_A$ *an-*] [$_N$ *aśva*]] are such compounds. An *upapada-tatpuruṣa* compound is one whose right sister is a bound, nominal morpheme derived from a verbal root. Examples of such bound morphemes are: *-bhid*, *-jña*, *-stha*, *-dṛś*, *-ghna*, *-cara*, etc.

On the assumption that one can identify a bound, nominal morpheme derived from the verbal root, one can easily identify a compound such as [$_N$ [$_N$ *sarva*] [$_N$ *-jña*]] as an *upapada-tatpuruṣa* compound. Finally, a *prādi-tatpuruṣa* compound is one whose first constituent is either a preposition (e.g., *pra*), a prefixing bound morpheme (e.g., *ku-*) or an indeclinable (e.g., *puras*). These morphemes are listed in the grammar with the first member of the list being the preposition *pra*. (Hence, they are given the name *prādi* compounds.) Each of these compounds, then, are readily identified from their analysis and the *prādi* list. Examples of such compounds are: [$_N$ [$_P$

[4] Since *aluk* compounds are not productive, they are listed in the dictionary and they can be listed with their analysis. Some questions of implementation arise with respect to whether or not one wishes to encode the case of the inflected subordinate word.

adhi] [$_N$ *rāja*]], [$_N$ [$_A$ *ku-*] [$_N$ *puruṣa*]] and [$_N$ [$_A$ *puras*] [$_N$ *-kāra*]].

We now come to compounds which require further annotation for their identification. They are those compounds comprising two adjectival stems, two nominal stems or a nominal stem followed by an adjectival stem.[5] Let us begin with stems of the form N N. *Dvandva* compounds, many *vibhakti-tatpuruṣa* compounds and *karmadhāraya* compounds are of this form. It is common to distinguish headed constituents from non-headed constituents. *Vibhakti-tatpuruṣa* compounds and most *karmadhāraya* compounds are headed, indeed, right headed. *Dvandva* compounds are non-headed compounds; and some *karmadhāraya* compounds are also non-headed.

The simplest extension of the notation is to introduce a special symbol for the non-headed compounds, inserting between the elements a plus sign, say. Thus, one would have the *dvandva* compound [$_N$ [$_N$ *rāma*]+[$_N$ *kṛṣṇa*]]. Some so-called *viśeṣaṇa-ubhaya-pada-karmadhāraya* compounds are also non-headed: for example, [$_A$ [$_A$ *snāta*]+[$_A$ *anulipta*]]. Now it is easy from this notation to see which is which. The compounds which have the plus sign and both of whose constituents are nouns are *dvandva* compounds; other compounds with the plus sign are *viśeṣaṇa-ubhaya-pada-karmadhāraya* compounds.

To distinguish among the remaining compounds, one could do the following: one could introduce a labelled relational symbol.[6] This symbol could be indexed by a numeral between one and seven. Each *vibhakti tatpuruṣa* would have the numeral corresponding to its case. Thus, one has [$_N$ [$_N$ *sukha*] $\leq_2$ [$_A$ *āpanna*]], [$_N$ [$_N$ *ākhu*] $\leq_3$ [$_A$ *daṁśita*]], [$_N$ [$_N$ *go*] $\leq_4$ [$_A$ *hita*]], [$_N$ [$_N$ *vṛka*] $\leq_5$ [$_A$ *bhīta*]], [$_N$ [$_N$ *rāja*] $\leq_6$ [$_N$ *puruṣa*]] and [$_N$ [$_N$ *īśvara*] $\leq_7$ [$_A$ *adhīna*]].

Those with the numeral one are identifiable as *karmadhāraya* compounds. In this way, the *viśeṣaṇa-pūrva-pada-karmadhāraya* compounds, as exemplified by the compound [$_N$ [$_A$ *dīrgha*] $\leq_1$ [$_N$ *kaṇṭha*]], the *viśeṣaṇa-ubhaya-pada-karmadhāraya* compounds, as exemplied by [$_A$ [$_A$ *tulya*] $\leq_1$ [$_A$ *śveta*]], the *upamāna-pūrva-pada-karmadhāraya*, as exemplified by [$_N$ [$_N$ *anala*] $\leq_1$ [$_N$ *uṣṇa*]], the *avadhāraṇa-pūrva-pada-karmadhāraya*, as exemplified by [$_N$ [$_N$ *rāja*] $\leq_1$ [$_N$ *ṛṣi*]], and the *upamāna-uttara-pada-karmadhāraya*,

[5] I am making the simplifying assumption that participial stems are labelled as adjectival stems.

[6] In earlier work, I used the symbol $\prec$.

as exemplified by $[_N\ [_N$ *puruṣa* $]\ \leq_1\ [_N$ *vyāghra* $]\ ]$ would all be identifiable as *karmadhāraya* compounds. I shall leave open the question as to how these subclasses might be distinguished from one another using the enriched context free notation advocated here.

Finally, we come to *bahuvrīhi* compounds. Such compounds are most easily identified by the fact that their final constituent is a noun but they behave like adjectives. When this is marked inflectionally, such compounds are easily identified. But this is not always so. For example, when a *bahuvrīhi* compound is the left sister of a compound or of a derivational suffix, it cannot be grammatically identified. Indeed, such cases are ambiguous; and one must rely on the context to figure out whether the compound is a *tatpuruṣa* or a *bahuvrīhi*. Moreover, if the last word of the *bahuvrīhi* compound is of the same gender as the noun it modifies, again it is ambiguous. The easiest way to annotate such compounds is with a phonetically null suffix (see Gillon [4] for discussion), but how precisely to implement that depends on how inflectional tagging is to be done, another complexity not addressed here.

5 Bound Morphemes

We have already seen some instances of the utility of the notion of a bound morpheme, to be formalized as subcategorization, in providing parses for Sanskrit compounds. I end the paper with an indication of still further uses. To begin with, consider the stems *pūrva*, *apara*, *adhara*, *uttara*, *ardha* and *madhya*. In compounds, they are adjectives, in phrases they are nouns. This distinction is nicely handled by treating the adjectives as bound morphemes and the nouns as unbound morphemes.

In addition, many words, which, when uncompounded, belong to one inflectional class, belong to another, typically the *a* stem inflectional class, when compounded. In each case, subcategorization can be used to handle the treatment of these stems.

1. The word *ṛc*, which has a consonantal stem, becomes the *a*-stem *ṛca*, when preceded by a word in a compound. The same holds for the words *pur* (*pura*), *ap* (*apa*), *dhur* (*dhura*) and *pathin* (*patha*).

2. The consonantal stem words *sāman* and *loman*, when preceded by the

prepositions *prati*, *anu* and *ava* become the *a*-stem words *sāma* and *loma*, respectively.

3. The *i* stem word *bhūmi* becomes the *a* stem word *bhūma*, when preceded in compound by either *kṛṣṇa*, *pāṇḍu* or a numeral.

4. The *ī* stem words *nadī* and *godavarī* become the *a* stem words *nada* and *godavara*, respectively, when preceded in compound by a numeral.

5. The *r* stem word *catur* becomes the *a* stem word *catura*, when preceded in compound by either *tri* or *upa*.

6. The *s* stem word *varcas* becomes *a* stem word *varcasa*, when preceded in compound by either *brahman*, *hastin*, *palya* or *rājan*.

7. The *s* stem word *tamas* becomes *a* stem word *tamasa*, when preceded in compound by *ava*, *sam* or *andha*.

8. The *n* stem word *adhvan* becomes *a* stem word *adhva*, when preceded in compound by a preposition.

9. The *i* stem word *aṅguli* becomes *a* stem word *aṅgula*, when preceded in compound by either a numeral or an indeclinable.

10. The *i* stem word *rātri* becomes *a* stem word *rātra*, when preceded in compound by either a numeral or an indeclinable or *ahan* or *sarva* or a word denoting part of the night.

11. The *n* stem word *ahan* becomes *a* stem word *aha*, when preceded in compound by either a numeral or an indeclinable or *sarva* or a word denoting part of the day.

6 Conclusion

Above, I have made a *prima facie* case that the information pertaining to the grammar of compounds in Classical Sanskrit captured in their classification by the Pāṇinian tradition can be retrieved from a very slight augmentation of the usual enriched context free rules used by generative linguists. I have reviewed this classification and I have shown, for each category in the classification, how that classification can be retrieved from a fairly standard set of enriched context free rules, adapted for Classical Sanskrit.

References

[1] Cardona, George 1988 *Pāṇini: his work and its traditions. Background and Introduction.* New Delhi: Motilal Banarsidass.

[2] Chomsky, Noam 1957 *Syntactic structures.* The Hague, The Netherlands: Mouton and Company (Janua Linguarum: Series Minor n. 4).

[3] Chomsky, Noam 1963 Formal properties of grammars. In: Luce, R. D.; Bush, R.; Galanter, E. (eds) 1963 v. 2, pp. 323–418.

[4] Gillon, Brendan S. 1995 Autonomy of word formation: evidence from Classical Sanskrit. *Indian Linguistics*: v. 56, n. 1–4, pp. 15–52.

[5] Kielhorn, Franz (ed) 1880 *The Vyākaraṇa Mahābhāṣya of Patañjali.* Poona, India: Bhandarkar Oriental Research Institute (4^{th} edition revised by R. N. Dandekar 1985).

[6] Luce, Robert D.; Bush, Robert; Galanter, Eugene (eds) 1963 *Handbook of mathematical psychology.* New York, New York: Wiley.

[7] Selkirk, Elizabeth O. 1982 *The syntax of words.* Cambridge, Massachusetts: The MIT Press.

[8] Whitney, William Dwight 1881 *Sanskrit grammar: including both the classical language, and the older dialects, of Veda and Brahmana.* Cambridge, Massachusetts: Harvard University Press, 2nd edition (1889), 11th reprint (1967).