Hoisting the colors of Sanskrit

Gérard Huet
Inria Paris Center

Abstract

The Sanskrit Heritage Reader is a tool that transforms a Sanskrit text into a representation of all its possible segmentations as a word-by-word enunciation (*padapāṭha*), undoing the sandhi euphonic rules. It allows a human annotator to select relevant segmentations with the use of a man-machine graphical interface using a notion of colored segment. We discuss in this paper this notion of color, and argue about its linguistic status, similar to a notion of parts of speech, but also allowing to segment complex compounds into a linear representation of so-called pre-compounds. This permits to present the *padapāṭha* with segmented compounds, without having to decide prematurely of their exact internal morphology. Colors can be seen as the coarsest classification of morphemes allowing the regular representation of Sanskrit. The study reveals subtle difficulties in Sanskrit analysis not usually discussed in grammars, which deal extensively with generativity (*vṛtti*), but rarely discuss text analysis (*śābdabodha*).

1 Introduction

The Sanskrit Heritage Reader[1] is a segmenting service for Classical Sanskrit. It takes as input a piece of Sanskrit text and proposes the various ways the text may be segmented into as a sequence of word forms (*pada*) by undoing phonetic glueing (*sandhi*). Such a sequence is called a word-by-word recitation (*padapāṭha*) of the piece of text (*śabda*). Actually, it does a bit more: it breaks compounds into their constituents, profiting of the fact that the intra-compound sandhi obeys the same rules as sentential sandhi. Thus the tool may be used for resolving compounds into constituents as well as providing word decomposition of sentences, in view of further analysis by various parsers to recognize its meaning.

This segmenter was designed as a finite-state relational process or Eilenberg machine (Huet, 2005). It assumes the prior generation of word forms and necessary phonemes from a given vocabulary defined by a generative lexicon, in our case the Sanskrit Heritage dictionary. The dictionary defines lexical items given with generation parameters, such as the present class (*gaṇa*) of a verb, or the gender of a nominal. Thus the various data-banks used by the segmenter store the corresponding signifier not just as a list of phonemes (*varṇa*) but as a tagged entity exhibiting its generation parameters. Its tags are used for informing the user with the generation parameters of the various pieces of the *padapāṭha*. Furthermore, the various kinds of segments are displayed with a characteristic colour, making explicit its combinatorial power. Thus finite verb forms such as *gacchati* are represented as red segments, nominal forms such as *mudrā* are represented as blue segments, and nominal stems like *hasta* are represented as yellow segments, usable as first component of compound form *hastamudrā*, itself represented as two consecutive segments, yellow *hasta* followed by blue *mudrā*.

This color-coding turned out to be very effective to select the right segmentation of a text from its possibly many potential segmentations. It is specially so when using the graphical interface

[1] https://sanskrit.inria.fr/DICO/reader.html

of the Reader (Goyal and Huet, 2016). The goal of this note is to give a systematic exposition of this notion of color, and to investigate its proper linguistic status.

2 Basic colors

The main partition of Sanskrit padas is between nominals, constructed with the *sup* declension suffixes, and verbal finite forms, constructed with the *tiṅ* conjugation suffixes, according to Pāṇini's sūtra (I,4,14): *suptiṅantam padam*. We use the color red for *tiṅanta* padas, this hot color being appropriate for the dynamic action that they denote, while the cool blue color is used for *subanta* nominal padas.

Thus on input *vavarṣarudhiram* we get the colored segments indicated in Figure 1.

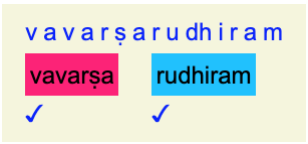

Figure 1: vavarṣarudhiram

When clicking in the interface on red *vavarṣa*, one gets its morphological tag as: `[vṛṣ]{pft. ac. sg. 3 | pft. ac. sg. 1}` and similarly on blue *rudhiram*, one gets `[rudhira]{m. sg. acc. | n. sg. acc. | n. sg. nom.}`. This shows that this segmentation may be interpreted in 6 different ways according to their inflexion *vibhakti*, and consequently as possibly various meanings. Here our colors are essentially marking part-of-speech of the words, no more.

Thus on sentence *bhīmorudhirampibati* we get the same blue color for the agent Bhīma and for the object of its drinking, blood, as shown in Figure 2.

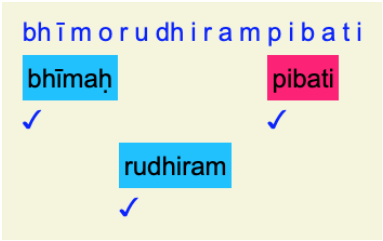

Figure 2: bhīmorudhirampibati

Actually, we set aside the vocative forms, colored green to make clear that they do not contribute to the meaning of the text, but rather serve are indications pertaining to the enunciation conditions. They belong to the discourse structure of the text, rather than to the sentence structure proper. Also, grammatical content-less words such as conjunctions and other adverbials, which are indeclinable stand-alone items, are colored mauve, to distinguish them from blue nominals, which denote participants (*kārakas*) having thematic role in the situation described by the input sentence. Finally, (inflected) pronouns are colored a lighter shade of blue than actual nominals.

Thus on input *deva adyamamabhāryāpacati* "Majesty, today my wife is cooking", we get the multicolor rendering shown in Figure 3.

3 Nominal Compounding

In Sanskrit, compounding is productive, without limitation on the number of its components. The standard compounding operation applies to two nominal forms, glueing by sandhi the stem of the first to the second one. Thus, a yellow cloth, *pītam ambaram*, may be contracted in one pada by compounding, yielding form *pītāmbaram*. We represent such a *karmadhāraya* compound

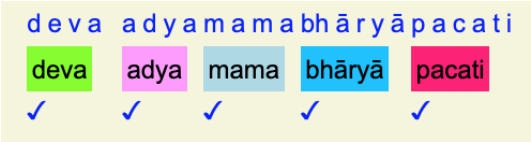

Figure 3: deva adyamamabhāryāpacati

with two segments, one yellow for the left stem *pīta*, preceding the blue right form *ambaram*, like in Figure 4 below.

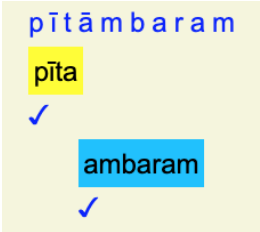

Figure 4: pītāmbaram

The notation generalizes in a straightforward manner to *dvandva* compounds, with possibly many yellow segments, like for *aśvāvyuṣṭrāḥ* (horses, sheep and camels), represented in Figure 5 below.

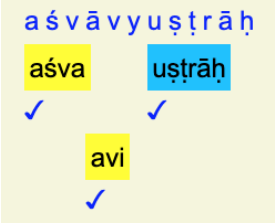

Figure 5: aśvāvyuṣṭrāḥ

But sometimes there may be an ambiguity, in case of an embedded compound such as *daśarathaputraḥ* which could be analysed theoretically as *(daśa-ratha)-putraḥ* or *daśa-(ratha-putraḥ)*. This ambiguity is often avoided by not segmenting proper names, when they are properly lexicalized, like here shown in Figure 6.

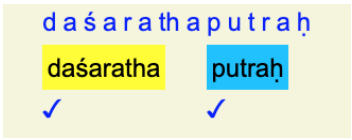

Figure 6: daśarathaputraḥ

Also, substantives obtained by compounding with a root form like *nṛpa*, king "who leads men" should be lexicalized as a frozen form (*rūḍha*). This also applies to lexical items like *malamūka* (deaf and dumb). This way most complex compounds may be understood un-ambiguously as left-associating, like in Pañcatantra's famous 10-components *samāsa* shown in Figure 7.

Now that we understand the representation of determinative (*tatpuruṣa*) compounds, let us move to the possessive class (*bahuvrīhi*). Here we have a problem, because such exocentric compounds, turning into adjectives, may transform their right hand component in order to assign the compound some gender that is not allowed for the original right-hand side nominal.

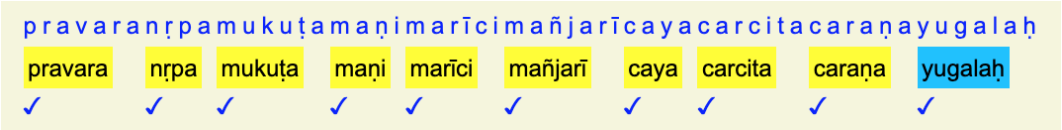

Figure 7: pravaranṛpamukuṭamaṇimarīcimañjarīcayacarcitacaraṇayugalaḥ

That is, their gender is not a synthesized attribute, but it is inherited from the surrounding noun phrase head. Thus we may get *pītāmbaraḥ* "who wears a yellow garment", typically as an epithet of Viṣṇu, where the masculine form *ambaraḥ* of its right component is not a valid form of the neuter substantive *ambaram*. We solve this problem by generating extra forms for such substantives, usable only in *ifc* (*in fine composi*) position. We assign the color cyan to such components, yielding the representation shown in Figure 8.

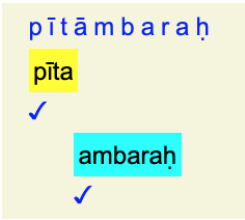

Figure 8: pītāmbaraḥ

This same mechanism is used to enter forms of lexical entries which are restricted to the ifc role, like *-kāra*, or root ifc forms. Thus for *kumbhakāraḥ* (pot-maker) we get Figure 9.

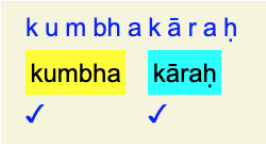

Figure 9: kumbhakāraḥ

An additional difficulty concerns feminine substantives like *daṃṣṭrā*. First, they must generate an iic (*in initio composi*) ended in the feminine suffix *-ā*, in order to allow recognition of a compound such as *daṃṣṭrākarālaḥ* "exhibiting horrific fangs". But we must also license a segment in the masculine stem *-daṃṣṭra-* to allow constructing a *bahuvrīhi* compound, itself usable as left component of further compounds. We use the khaki color for this (rather rare) occurrence. Here is an example, for *bhagnanakhadaṃṣṭravyālam* "wild beasts deprived of their claws and fangs", shown in Figure 10.

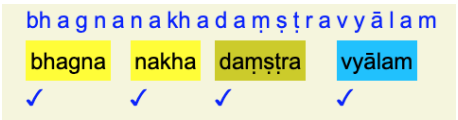

Figure 10: bhagnanakhadaṃṣṭravyālam

Please note that such khaki segments signal their use both as an ifc on their left, to build a stem now used as iic to their right. Thus it is a stem that is not an iic, but rather the last component of the *prātipadika* of a compound ended in an ifc component. Thus in *kumbhakāra-putraḥ* segment *kumbha* is yellow, but segment *kāra* is khaki, as right component of the *prātipadika* of compound pada *kumbhakārasya*. It thus partly disambiguates the grouping of stems used to build a complex compound.

Most regular compounds are left-associative, like the example in Figure 7. The successive compositions of meanings allows understanding such long compounds in real time, without taxing short-term memory. Many of the exceptions to this rule concern proper names, and frozen technical terms which may not be compositional in meaning. Such items should be lexicalised, and the compounds containing them will be linear, understandable by left to right composing of the meanings of its individual segments, like in the above example of Figure 7.

At this point we note that, even if proper names are lexicalized, we must allow for the possibility of using a descriptive compound in the vocative. For instance, on input *namaste'stu mahāmāye śrīpīṭhe surapūjite* (I honor Thou, Great Illusion, Enthroned by Fortune, Blessed by the Gods), we get Figure 11.

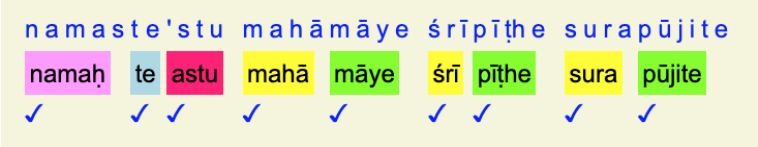

Figure 11: namaste'stu mahāmāye śrīpīṭhe surapūjite

This concludes our discussion of regular nominal compounds. There exist also exceptions to the two rules that are implicit in those: firstly, the first component does not carry the *vibhakti* suffixes, and is reduced by *luk* to its stem form; secondly, retroflexion does not cross over the compound frontier. When one of these two conditions is not met, the compound must be lexicalized. For instance, consider *agrevaṇam* "border of the wood", where the first component is in the locative (*aluk*), and the second component *vanam* incurs retroflexion because of its left context. Such compounds are in small number, mostly to form proper names like *Janamejaya* or *Rāmāyaṇa*.

4 Adverbial compounds

An important, although diverse, family of compounds is referred to by the name *avyayībhāva* "turned into an indeclinable". A typical representative is *yathāvṛddhi* "according to the phase of the moon". Its left-hand side is the invariable *yathā*, and its right side is the neuter form *vṛddhi*. Here this right component cannot be used stand-alone, since it is not a form of feminine nominal *vṛddhi*. We chose to represent such a compound with 2 segments, a pink iic and a mauve ifc, as shown in Figure 12.

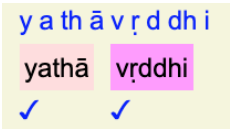

Figure 12: yathāvṛddhi

The *avyayībhāva* compound family comprises many unusual constructions, and sometimes the indeclinable is in its right hand part rather than its left part, like *sūpaprati* "with some sauce". Such exceptional items must be lexicalized.

Absolutives are colored mauve, in keeping with their adverbial nature, but a specific color could have been specified. Absolutives in *-tvā* are provided for roots without preverbs, absolutives in *-ya* attach to verbs provided by preverbs, which are represented glued to their root, like finite verbal forms. They are thus all mono-segment. Occasionally they appear negated, in which case we get two segments, the privative prefix being a stand-alone segment, like in the following example, *yato vāco nivartante aprāpya manasā saha* "*(brahman* is) where speech returns from, unable to grasp it with the mind", shown in Figure 13.

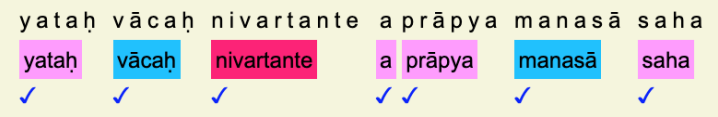

Figure 13: yato vāco nivartante aprāpya manasāsaha

Also in mauve are occasional absolutives in *-am* (the so-called *ṇamul* construction), and infinitive forms in *-tum*. At this point it should be mentioned that the various participial forms are colored blue, in keeping with their nominal declension.

Other generative adverbs are the so-called *tasil*, adverbs of manner like *ubhayatas* (from both sides), colored mauve.

5 Verbal compounds

Finite forms of verbs are usually represented as a single segment. That is, potential preverbs are glued to the root form, contrarily to the original design of our Reader, where preverbs were represented as separate segments. Similarly, prefixes such as prepositions, but also particles such as *sa-*, *su-*, *dus-*, *ku-*, etc are considered morphemes of the inner morphology of padas, and not segments to be compounded. This also applies to the privative prefixes *a-*/*an-*. This induces extra lexicalisation, but greatly simplifies the user interface. This is also the case for secondary *taddhita* suffixes such as *-tva*, *-tā*, *-vat*, etc.

There is an exception for the so-called periphrastic perfect, obtained by glueing a special morpheme in *-ām* to a form in the perfect of one of the auxiliary verbs *as*, *bhū* and *kṛ*. We represent such forms as two segments, the form in *-ām* in orange, while the auxiliary perfect form is a usual verbal form in red; Thus, for sentence *āsāṃcakre hvayāmāsa ca* (he sat and called) we get Figure 14. By contrast, so-called periphrastic future forms are monosegmental, treated as a specific conjugation.

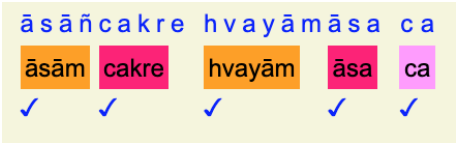

Figure 14: āsāṃcakre hvayāmāsa ca

Another kind of verbal compound is the so-called *cvi* inchoative construction. It allows root forms of the 3 auxiliaries to be prefixed with nominal stems in *-ī* (or sometimes *-ū*). Some other initial segments like *sākṣāt* may also be used, the class of such segments is called *gati*. By analogy with periphrastic perfect, we use the orange color to represent such initial segments. Finally, the construction is extended to nominal compounds, formed with a *gati* followed by a declined *kṛdanta* (first level nominal derivative) of the auxiliaries. Figure 15 shows a few typical such forms.

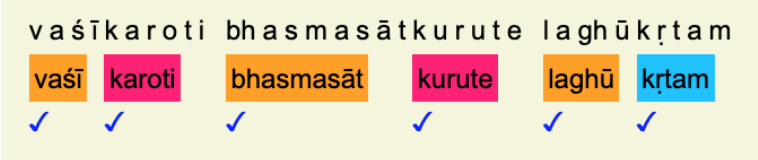

Figure 15: Various inchoative compounds

It should be noted that these nominal inchoative compounds may themselves be subject to compounding, like in *vivādāspadībhūtam* (that has become litigious) in Figure 16 below. Please

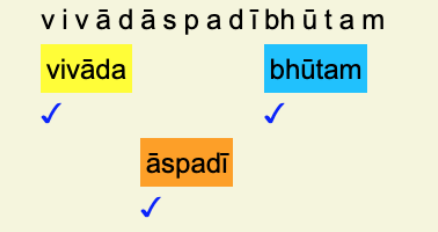

Figure 16: vivādāspadībhūtam

note that we have here one pada represented as three segments.

6 Special constructions

One last productive morphological construction is exemplified by the bahuvrīhi compound *vaktukāmaḥ* "desirous of speaking". Its left component is an infinitival form deprived of its final *-m*, and its right component is a form of *kāma*, *manas* or *śakya*. This construction is not very frequent, but it is productive, used by the best authors, and stated in a *vārttika* to *sūtra* (VI,1,144). It must therefore be accommodated, as a separate construct. We use again the orange color for the infinitive segment *vaktu*, in analogy with the verbal compounds above, while we us the cyan color for the ifc segment *kāmaḥ*. We show in Figure 17 the example *punarapi vaktukāma ivāryo lakṣyate* "Your honor appears desirous of speaking again".

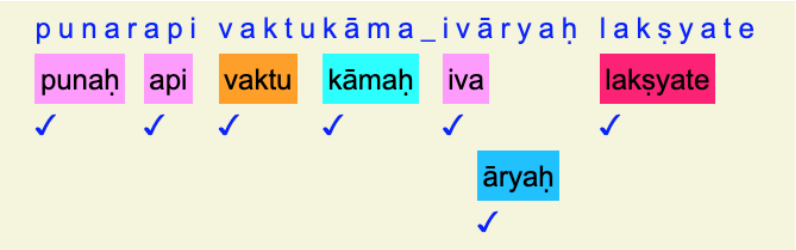

Figure 17: punarapi vaktukāma ivāryo lakṣyate

It is not always obvious to decide whereas an unusual formation ought to be productive or treated as a lexicalized exception. For instance, *sūtra* (II,1,72) alludes to irregular compounds in the group (*gaṇa*) headed with *mayūravyaṃsaka*. This group is indeed a motley crew, with bizarre isolated items like *jahijoḍaḥ* (in the habit of hitting one's chin) which may be simply lexicalized. Others mentioned in *vārttika*s allude to generic schemas like *aśnītapibatā* built on two imperative forms compounded as a feminine substantive, meaning a festive occasion where the two actions are repeatedly performed, like in this case eating and drinking. Indeed the *gaṇapāṭha* lists 12 such attested substantives, raising the question of whether this construction is productive. We decided against accepting this scheme as generic, because this would oblige us to introduce two specific colors, one for imperative iic segments, the other for imperative ifc segments construed as feminine substantives, and we rather decided to lexicalize the 12 attested occurrences. One may even question whether such forms, completely atypical in making nominal compounds out of verbal forms, deserve the status of grammatically correct Sanskrit. Would such crude colloquial expressions have been considered *śiṣṭa* by Pāṇini? Aren't they just tongue-in-cheek suggestions by Kātyāyana as implicit mockery of grammar's pretension to straight-jacket a living language?

We end this section by mentioning one last color, grey, for input chunks that are not recognized by the lexer. This may be because of an incompleteness of the generative lexicon, or this may be due to an un-grammatical item in the input, like shown in Figure 18, for input *na tathā bādhate śītaṃ yathā bādhati bādhate* "It's not so much the cold as 'bādhati' that bothers me", told by a palanquin bearer to the bogus pandit he is carrying, who had asked: *api śītaṃ bādhati*?

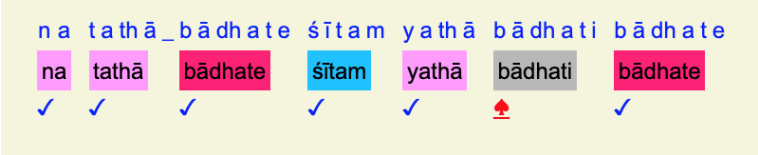

Figure 18: na tathā bādhate śītaṃ yathā bādhati bādhate

7 Making linguistic sense of colors

We have given a fairly extensive treatment of the visualisations of various Sanskrit constructions, as they are recognized by the Sanskrit Heritage segmenter, and as they are displayed in the graphic interface of its Reader. In this visualisation, various segments are displayed in various colors giving some information about morphemes that are necessary to represent Sanskrit padas. Most of the complexity of the treatment is due to the numerous ways of formation of compounding, and specially the changes of gender and number suffixes that occur along compound derivation. However, it is not straightforward to give a precise characterisation of our color encodings in terms of linguistic concepts.

Let us first sum up the various color encodings in the following table:

color	*part of speech*
blue	inflected nominal
red	finite verb form
yellow	nominal stem as iic
green	vocative
mauve	indeclinable, particle
light blue	pronoun
cyan	ifc
khaki	ifc stem
pink	indeclinable iic
orange	perfect iic, gati, infinitive stem
grey	unrecognized

For one thing, the colors presented above are ambiguous. For instance, the orange color is assigned to three kinds of morphemes: periphrastic perfect stems in *-ām*, *gati* verbal prefixes, and infinitives deprived of their final *-m*. It is obvious that these three constructions are independent, and that the common color is just merging these three constructions into a general category of "verbal compounds". We could indeed easily distinguish the three by visualising them differently, and each shade of orange would then designate unambiguously a precise grammatical operation.

Then, the status of "segment" itself is not entirely obvious. Actually, each segment generally represents a choice among several possible interpretations, explicit only if we click on the segment to uncover its possibly many tags. Consider for instance *pītāmbaraḥ* in Figure 9 above. The cyan component *ambaraḥ* is tagged {`m. sg. nom.`}. This does not mean that *ambaraḥ* is a valid masculine substantival form, which it is not, just that the particular segmentation *pīta-ambaraḥ* may be interpreted as the masculine form of an adjective obtained by the bahuvrīhi interpretation of the compound: "he who has a yellow garment", in the nominative case. Thus the tag is not a tag of the segment, but of the whole compound pada. Now consider *pītāmbaram*. Here *ambaram* may have two colors. As a cyan segment, it would be tagged {`m. sg. acc.`}, meaning similarly that this particular segmentation *pīta-ambaram* may be interpreted as the same masculine adjective, but now in the accusative case. Whereas if it appears as a blue segment, it bears the multitag {`n. sg. acc. | n. sg. nom.`}, meaning that here *pīta-ambaram* could be interpreted either as the *karmadhāraya* compound obtained by contraction of *pītam ambaram* "yellow cloth" (in the nominative or accusative case), or possibly the bahuvrīhi

adjective, but now fit with the neuter gender, in order to serve as determinant to some neuter head noun. This shows that in general, even if we restrict segments to single tags, a given compound segmentation does not characterize its mode of formation.

We must explain the reason for this apparent confusion. It stems from the morphological generation process, that compiles the lexicon entries by feeding the various databanks corresponding to our colors. On processing the entry *ambara*, which is marked as a neuter substantive, it generates ifc items (cyan) only for the missing genders, in order to be able to recognize masculine and feminine instances of the bahuvrīhi adjective. Indeed, on entry *pītāmbarā*, our segmenter generates only a cyan ifc *ambarā*, with tag {`f. sg. nom.`}. The rationale of not generating a cyan *ambara* of gender neuter when analysing *pītāmbaram* was that it would just be redundant with the *karmadhāraya* segmentation, from the point of view of possible tags. Thus it is left to the next stage of interpretation after segmentation, i.e parsing, to make sense of sequences of padas decorated with *vibhakti* parameters. And it is the parser that must guess, in case of the neuter gender, whether the segment fits with the surrounding context, in order to build the agreeing nominal phrases. Thus in the case of a neuter interpretation, it is the burden of the parser to choose between the two interpretations of the compound pada as a substantive (*viśeṣya*) or as an adjective (*viśeṣaṇa*). Whereas in the case of cyan segments, there is no such choice, they must be adjectives. Unfortunately, the parser does not have the information, if we transmit to it just morphological tags, but not the color that in the cyan case bears information.

This discussion shows that there is not a good transmission of information if the result of the segmenter is just a list of morphological tagged pada. It suggests that we could do better, if the *viśeṣya/viśeṣaṇa* status of compounds is transmitted when known. And this is where colors may help. But in order to deal correctly with the neuter case, we must distribute it in the two colors. Thus on *pītam ambaram* we could generate a blue *ambaram* bearing tags {`n. sg. acc. | n. sg. nom.`}, as presently, but also a cyan *ambaram* bearing now tags {`n. sg. acc. | n. sg. nom. | m. sg. acc.`}. So we would have more singly tagged segmentations, but now we can transmit their substantive/adjective character with the color, that would have now a clearer linguistic status. Thus we could generate more precise information by transmitting the color along with the tagging, but at the price of overgeneration — more solutions would be allowed.

Unfortunately, there are many more compound constructions which are ambiguous with regard to their *viśeṣya/viśeṣaṇa* status, specially when compounding is iterated, typically in poetic style (*kāvya*). Thus when we recognize a series of embedded compounds $X_1X_2...X_nY$ we only propose a list of n yellow stems (*prātipādika*) followed by a blue (or cyan) pada, and this ignores not only the exponentially many ways in which the X sequence is the frontier of binary compounds, but also the *viśeṣya/viśeṣaṇa* status of internal binary compounds, according to their *bahuvrīhi/tatpuruṣa* actual construction. Not to speak of the fact that a consecutive sequence of X's and possibly Y might represent one dvanda multi-component compound.

Two remarks are in order. The notation of syntactic constituents used by Gillon to represent phrase-structure in Sanskrit has a special mark "-B" to indicate *bahuvrīhi* raising (Gillon, 1995; Gillon, 2009). Unfortunately, this mark has a null phonetic realization ("morphological zero") and this is one of the main causes of Sanskrit ambiguity. Sometimes, Navyanyāya terms, using compounding for relational expressions, occasionally use suffixing by a taddhita pratyaya *-ka* (technically called *kaP*) to make explicit *bahuvrīhi* raising. In this case, the ambiguity is lifted by explicit phonetic marking — the extra *ka* syllable is the information that decides its *viśeṣaṇa* status. Another device is the accent. Accent on the first component marks its *bahuvrīhi* character. Unfortunately, accent is not marked in Classical Sanskrit, so the parser must decide the matter. And actually, in poetry, the ambiguity may be used to have two different interpretations in the two branches of a double-entendre (*śleṣa*).

This discussion shows that it would be wrong to interpret the two colors blue and cyan as part-of-speech markers, identifying respectively substantives and adjectives. They both pertain

to nominals, i.e. subantas, the *viśeṣya/viśeṣaṇa* status of which is not intrinsic, but depends on the context. Furthermore, the distinction between nouns and adjectives in Sanskrit is not completely clear. Actually, even an authority such as S. D. Joshi says: 'it is very difficult to provide a satisfactory definition of the concepts *viśeṣaṇa* and *viśeṣya*, because we don't have adequate criteria for the differentiation of adjectives and substantives"(Joshi, 1966). Additional discussion on this question is provided in (Dash, 1987).

Actually, the design rationale of the segmenter is just to build the *padapāṭha*, and to transmit to the next layer of interpretation the extra information it may have gathered on the way; thus, the Heritage segmenter, being lexicon-directed, is aware of morphological tags, which are parameters of their synthesised forms, generated from the lexicon, and thus transmits padas with their *vibhakti*, but should not try to guess thematic roles (*kāraka*) or even recognize noun phrases. It is up to the next stage of parsing to understand the situation semantics. Thus, the Heritage shallow parser (Huet, 2007), just by grouping phrases in the first three cases (nominative, accusative, instrumental), guesses the thematic roles of the situation, without having to understand the structure of compounds. In comparison, Amba Kulkarni's dependency parser (Kulkarni, 2019) goes further in analysing sentences, and uses semantic principles such as compatibility (*yogyatā*). This way, her parser is able to go as far as translation to Hindi.

8 Geometry of colors

Our colors actually indicate states of the segmenting automaton, which correspond to a mixture of morphological categories and their mutual combinatorics. Each color corresponds to a databank of morphemes. These morphemes are assembled by external sandhi in order to form padas, according to the finite automaton graph shown in Figure 19 below.

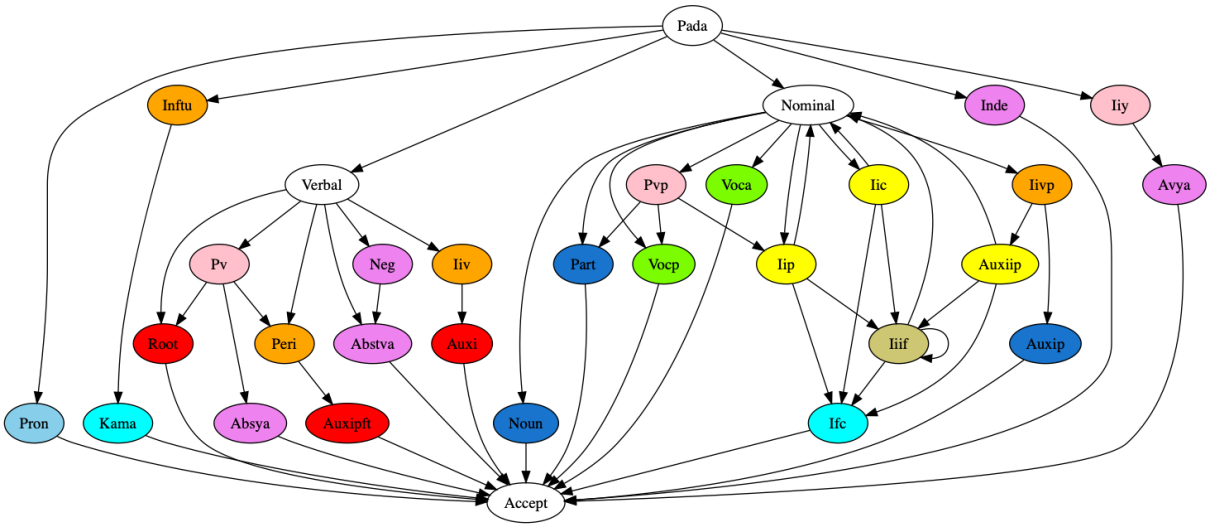

Figure 19: Colors geometry of padas

Our design of the segmenter evolved over time. Initially, preverbs were recognized as independent segments. But this feature was deemed confusing by users, with non-connected preverbs floating around. Thus we decided to glue them on the fly, either to finite verb forms of the roots, or to their participial forms (*kṛdanta*). At glueing time, we recognize that the particular sequence of preverbs is consistent with the roots' requirements. This allows us to dispense with huge databanks storing the glued preverbs, and to satisfactorily deal with problematic sandhi operations like in the recognition of *ihehi*, with a mechanism of *phantom phonemes* (Huet, 2006). It is to be noted that assuming preverbs as prefixes of verbal forms may prevent recognition of some Vedic sentences, where prepositions may float around more freely.

Thus preverbs still appear in the finite automaton state transition diagram, but are silently

glued to verbal forms when vizualized by the reader. This is important for instance for properly recognizing absolutive forms, ending in *-tvā* for roots, and in *-ya* for verbs prefixed with preverbs (using *kṛt* suffixes respectively *Ktvā* and *lyaP*). This induced us to generate *kṛdanta* forms, notably participial forms, in a separate databank from other nominals (see Part state in Figure 19). Such forms are generated with compound tags indicating their internal morphology. For instance, *pragatam* is tagged as `[pra-gata { pp. }[pra-gam]]{n. sg. acc. | n. sg. nom. | m. sg. acc.}`. This information is useful for further stages of parsing, since they are useful to determine their combinatorial power (*ākāṅkṣā*). However, we kept the blue colors for such *kṛdanta* forms, in an effort to avoid confusion of the annotator. We could of course have distinguished them with some different shade of blue, but the information is there anyway in the tag when needed.

Another improvement was to treat privative prefixes *a-*/*an-* as part of internal morphology, as opposed to independent iic. segments. This was crucial for helping annotators to notice situations of ambiguity where a segment ending in *-ā* precedes a possibly negated notion, leading to two opposite analyses. On the other hand, this obliged us to lexicalize such privative compounds, and this is problematic for long-scope negation, where the privative prefix operates on a compound (Gillon, 1987). This treatment of negation is homogeneous with the treatment of nominal prefixes like *su-*, *dus-*, *sa-*, vi, etc. This relieves the segmenter from overgenerating with such mono-syllabic particles if treated as genuine stems, but at the cost of having to lexicalize the corresponding compounds. Symmetrically, we treat taddhita suffixes similarly. Again, this is problematic when these particles have a long scope, affixed to possibly long compounds. This also prevents recognition of long compounds used in Navyanyāya relational terms, where taddhita suffixes may cascade arbitrarily, often in un-Pāṇinian ways.

9 Psychology of colors

From the point of view of programming, our colors are implemented as a discrete set of atoms, that does not have any structure. However, from the point of the human-machine interface, our colors are mapped into RGB color representations, which are carefully chosen tints with features like hotness/coldness along a scale of activity associated with the corresponding segment of speech. Thus verbal forms, denoting actions, are bright red, while substantives are painted of a deep sky blue cold color since they have a more static role as participants to a situation. Pronouns are like budding nouns, of a ligher shade of blue. Exocentric [*bahuvrīhi*] compounds, of cyan color, as adjectives show some eagerness to associate with a deeper blue substantive they will qualify. Indeclinables are mauve, showing some intermediate status between nouns and verbs. Finally vocatives are of a flashy acid green, urging the user to either discard them, or to admit them when appropriate, but as discourse operators independent of the phrasal structure. Compounds have their iic marked as yellow, orange, or pink, according to the nature of their ifc, and with a slightly paler tint.

Thus these color codings have some rationale as psychological conceptual triggers, and this is very important from the point of view of helping the user to select the appropriate segmentation solution. Some care was taken also to have a consistent palette of colours, not too aggressive, and combining smoothly in order to offer a satisfactory esthetic experience. It is to be remarked that Amba Kulkarni's Saṃsādhanī platform uses also colors in her interface, and these colors are reminiscent of our own encodings, but with more shades of colors, because her parser interface needs to distinguish clearly the cases of nominals, in order to group them easily as noun phrases by agreement of their respective colors. In our segmenter, we made the choice of having few colors (10 in total) in order not to have the annotator lose time in a finer grain selection that is not relevant at segmentation time.

10 Conclusion

We have given a complete description of a notion of color which characterizes sorts of Sanskrit morphemes necessary to account for the notion of pada. All productive constructions of surface Sanskrit morphology are covered. We may consider this notion of color as a notion of parts of speech appropriate for Sanskrit. Segment colors are not independent, since the color transitions are governed by the Sanskrit segmenting finite-state automaton when recognizing padas. Thus, our colors indicate parts of speech categories, but now obeying algebraico/geometric laws corresponding to their mutual interactions as morphemes in pada production, and obtained by quotienting the state space of the segmenting automaton.

In this study, we have extended the initial dichotomy between subantas and tiṅantas in order to accommodate the various notions of compound word. The notion of nominal stem (*prātipādika*) came prominent, a sequential list of them being the key to collapse long compounds in the linear frontier of their parse trees, gaining an exponential factor by avoiding premature settlement of non-deterministic choices that are best solvable in the subsequent workflow of the global parser. This puts in evidence the need to introduce a variety of morpheme categories and their mutual connectivity requirements. It is satisfactory that most of these categories correspond to Pāṇinian notions, such as *cvi*, *gati*, *tasil*. Even less expected notions, such as the (khaki) stems of ifc-only items used as pseudo-iics, are known in the tradition as *bhāṣitapuṁska*. This conversion to the masculine stem of iics is the basis for our generation of yellow iic segments, but we need it also for conversion to masculine stem of inner compounds built on an item usable only in ifc, a subtle consideration.

One construct stands out as exceptional, allowing mixture of tiṅanta material (the infinitive *tumun* with *lopa* elision of final *-m*) with subanta padas (forms of *kāma*, *manas* and rarely *śakya*). This construct is often passed over in Sanskrit grammars, or stated as belonging to the late language (Whitney, 1924)§968g, (Apte, 1885)§181, (Renou, 1956) p72, but it is productive in classical Sanskrit. Indeed it is not discussed in Pāṇini's Aṣṭādhyāyī, and only appears in Kātyāyana's vārttikas to sūtra (VI,1,144); The same goes true for perfect periphrastic forms, which are exceptional in Vedic, restricted to auxiliary *kṛ* in sūtra (II,1,40), and only described fully in Patañjali's Mahābhāṣya (Renou, 1956) p72. We can thus witness diachronic evolution of the language by the necessary successive adjunctions of items to Figure 19. Refer also to the discussion on *aśnītapibatā* above.

The notion of pre-compound, allowing the collapse of an arbitrary number of embedded compounds, has put in evidence the necessity of providing segments which are stems of forms which normally occur only as right-hand sides of compounds, exemplified by the khaki *daṣṭra* in Figure 10. This problem is not discussed in Western Sanskrit grammars, to our knowledge. Such rare segments serve to partly disambiguate the compounds in which they appear, since they mark a boundary between two different compounds.

We have also departed from the standard presentations by separating vocative forms, deemed to belong to discourse analysis rather than sentential recognition. Such vocatives (as well as interjective particles like *bhoḥ*) are colored as acid green, and stand out in the interface in order for the annotator to decide between them and iic stems, since this is a frequent ambiguity, notably for stems in *-a*. The rationale is that the choice of the vocative must be consistent with the global context, known to the annotator from the particular text she/he is studying, or from semiotics considerations.

In summary, we have proposed a regular description formalism for *padapāṭha* sequences based on a notion of colored segment, deemed to be complete for Classical Sanskrit, and appropriate for quick selective annotation of corpus by human users with minimum training.

References

Vāman Shivarām Apte. 1885. *The Student's Guide to Sanskrit Composition. A Treatise on Sanskrit Syntax for Use of Schools and Colleges.* Lokasamgraha Press, Poona, India.

G. Cardona. 1988. *Pāṇini: his work and its traditions.* Motilal Barnasidass.

Siniruddha Dash. 1987. Adjectives and substantives as separate categories in Sanskrit. *Lokaprajñā*, 1,1:90–96.

Pierre-Sylvain Filliozat. 1988. *Grammaire sanskrite Pâninéenne.* Picard, Paris.

Brendan S. Gillon. 1987. Two forms of negation in Sanskrit: *prasajyapratisedha* and *paryudāsapratiṣedha*. *Lokaprajñā*, 1,1:81–89.

Brendan S. Gillon. 1995. Autonomy of word formation: evidence from Classical Sanskrit. *Indian Linguistics, 56 (1-4)*, pages 15–52.

Brendan S. Gillon. 2007. Exocentric (bahuvrīhi) compounds in classical Sanskrit. In Gérard Huet and Amba Kulkarni, editors, *Proceedings, First International Symposium on Sanskrit Computational Linguistics*, pages 1–12.

Brendan S. Gillon. 2009. Tagging classical Sanskrit compounds. In Amba Kulkarni and Gérard Huet, editors, *Sanskrit Computational Linguistics 3*, pages 98–105. Springer-Verlag LNAI 5406.

Pawan Goyal and Gérard Huet. 2016. Design and analysis of a lean interface for Sanskrit corpus annotation. *Journal of Linguistic Modeling*, 4(2):117–126.

Gérard Huet. 2005. A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *J. Functional Programming*, 15,4:573–614.

Gérard Huet, 2006. *Themes and Tasks in Old and Middle Indo-Aryan Linguistics, Eds. Bertil Tikkanen and Heinrich Hettrich*, chapter Lexicon-directed Segmentation and Tagging of Sanskrit, pages 307–325. Motilal Banarsidass, Delhi.

Gérard Huet. 2007. Shallow syntax analysis in Sanskrit guided by semantic nets constraints. In *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*, New York, NY, USA. ACM.

S. D. Joshi. 1966. Adjectives and substantives as a single class in the parts of speech. *Publications of the Centre of advanced study in Sanskrit, University of Poona*, Class A, no 9.

Amba Kulkarni. 2019. *Sanskrit parsing based on the theories of Śābdabodha.* Indian Institute of Advanced Study (DK Printworld distr.).

K V Ramakrishnamacharyulu Pavankumar Satuluri and Amba Kulkarni. 2016-17. Order of operations in the formation of Sanskrit compounds with special reference to introduction of samāsānta element and deletion of case endings. *Journal of Oriental Institute, Vadodara*, 66(4):77–86.

Louis Renou. 1942. *Terminologie grammaticale du sanskrit.* Honoré Champion, Paris.

Louis Renou. 1956. *Histoire de la langue sanskrite.* Editions IAC, Lyon.

Louis Renou. 1966. *La Grammaire de Pāṇini.* Ecole Française d'Extrême-Orient, Paris.

Pavankumar Satuluri and Amba Kulkarni. 2013. Generation of Sanskrit compounds. In *Proceedings of ICON 2013, the 10th International Conference on NLP*, pages 77–86, Noida, India.

Rama Nath Sharma. 1987-2003. *The Aṣṭādhyāyī of Pāṇini (6 vols).* Munshiram Manoharlal Publishers.

William Dwight Whitney. 1924. *Sanskrit Grammar.* Leipzig. 5th edition.