Appropriate Dependency Tagset for Sanskrit Analysis and Generation

Amba Kulkarni
University of Hyderabad

ambapradeep@gmail.com

April 28th, 2019

Natural Language Processing (NLP) aims to use computers to reduce the language barrier

- among humans speaking different languages,
- between a human being and a machine.

This involves two processes

- analysis, and
- generation.

Analyser: Analyses a surface string into a semantic structure

Generator: Generates a surface string from the underlying semantic structure

Semantic structure consists of relations between words

Analysis: Representation

dṛṣṭvā tu pāṇḍavānīkam vyūḍham duryodhanas tadā |
ācāryam upasaṅgamya rājā vacanam abravīt ||(BhG 1.2)

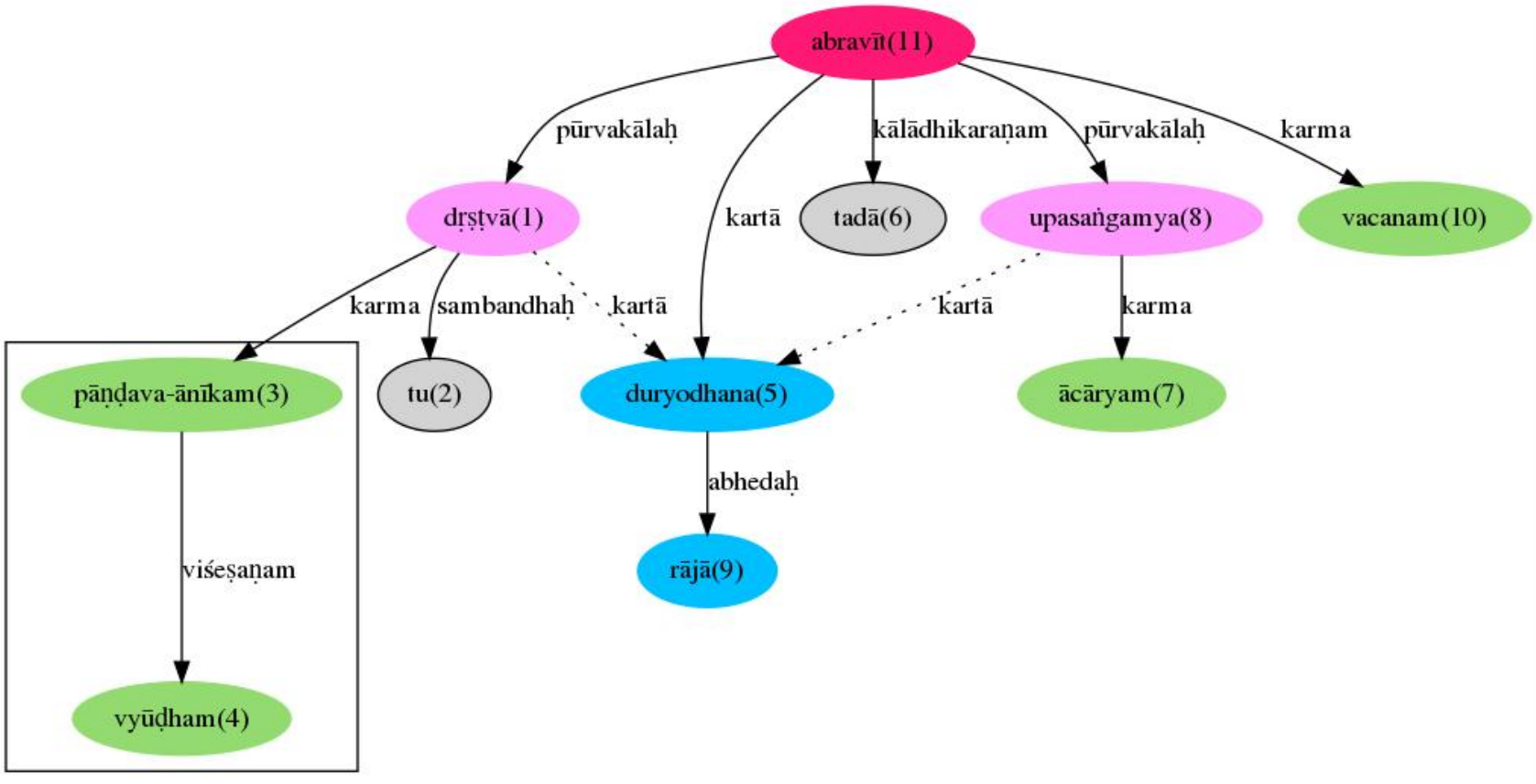


Figure: Analysis produced by the parser for BhG 1.2

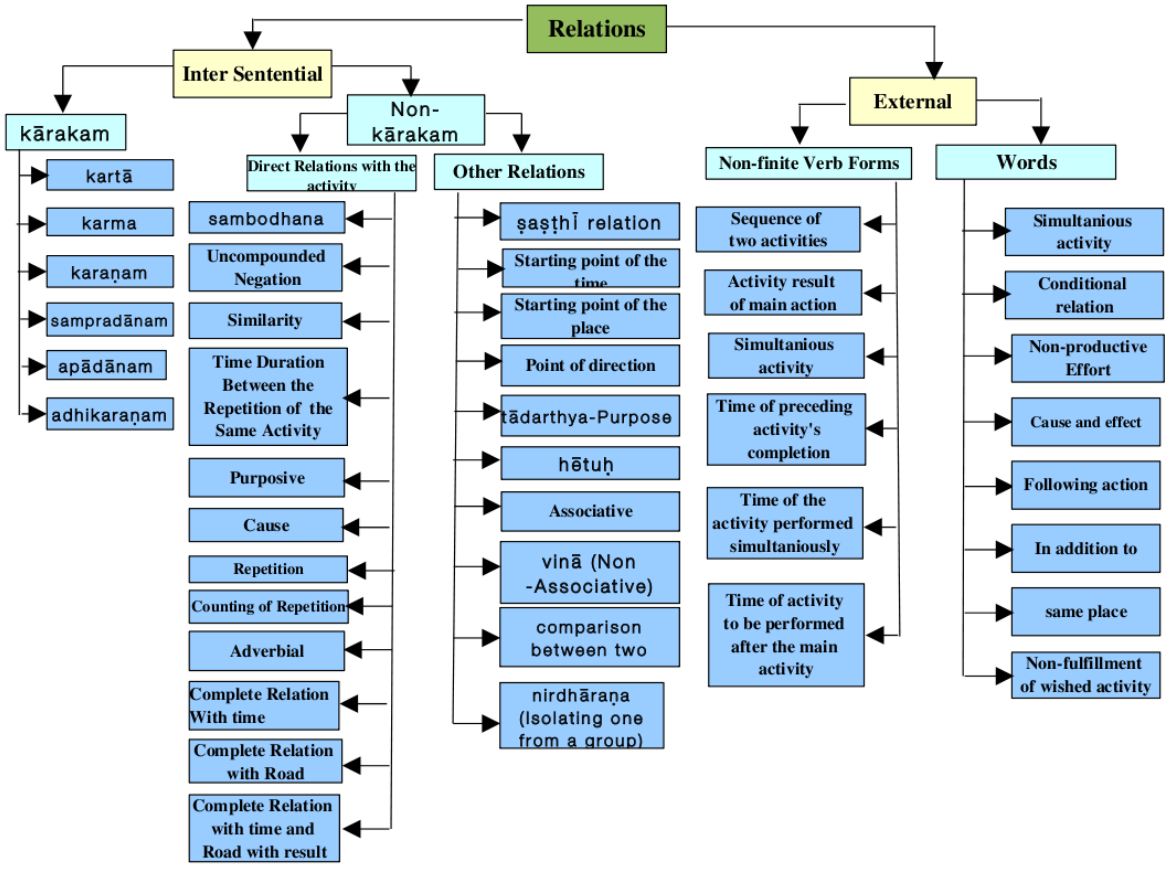
word index	stem and features	relation
1	dṛś1	pūrvakālaḥ 11
2	tu	sambandhaḥ 1
3	pāṇḍava-anīka {puṃ eka}	karma 1
4	vi+vah1 {bhūtakarma}	viśeṣaṇam 3
5	duryodhana {puṃ eka}	kartā 11
6	tadā	kālādhikaraṇam 11
7	ācārya {puṃ eka}	karma 8
8	upa_sam+gam1	pūrvakālaḥ 11
9	rājan	upādhi 5
10	vacana {napuṃ eka}	karma 11
11	brū1 {kartari anadyatanabhūtaḥ}	

Table: Machine internal representation of Analysis

Pāṇini uses several relation denoting terms

Theories of śābdabodha further provide fine-grain classification of such relations.

A set of such relations was compiled by K V Ramakrishnamacharyulu.



The granularity of these relations was too fine for mechanical processing.
For example, kartā was sub-classified into

- kartā
- karma-kartā
- karaṇa-kartā
- anubhavī-kartā
- prayojaka-kartā
- prayojya-kartā
- madhyastha-kartā
- amūrta-kartā
- utpreraka-kartā
- śaṣṭhī-kartā

The consortium for Sanskrit-Hindi Machine Translation adapted a subset of these relations for sentence level annotation of Sanskrit texts.

Granularity was increased and a proper subset is chosen that was appropriate for computational parsing.

kartā	prayojakakartā	prayojyakartā
karma	gauṇakarma	mukhyakarma
karaṇam	sampradānam	apādānam
adhikaraṇam		
deśādhikaraṇam	kālādhikaraṇam	viṣayādhikaraṇam
kartṛ-samā	karma-samā	kriyāviśeṣaṇam
hetuḥ	prayojanam	niṣedhyaḥ
pūrvakālaḥ	samānakālaḥ	anantarakālaḥ
sambodhanam	sambodhyaḥ	nirdhāraṇam
tādarthya	viśeṣaṇam	**ṣaṣṭhīsambandhaḥ**
vīpsā	pratiyogī	anuyogī
samuccitam	**upapadasambandhaḥ**	sambandhaḥ

Can we use the same set of relations for generation ?

word index	stem and features	relation
1	dṛś1	pūrvakālaḥ 11
2	tu	sambandhaḥ 1
3	pāṇḍava-anīka {puṃ eka}	karma 1
4	vi+vah1 {bhūtakarma}	viśeṣaṇam 3
5	duryodhana {puṃ eka}	kartā 11
6	tadā	kālādhikaraṇam 11
7	ācārya {puṃ eka}	karma 8
8	upa_sam+gam1	pūrvakālaḥ 11
9	rājan	upādhi 5
10	vacana {napuṃ eka}	karma 11
11	brū1 {kartari anadyatanabhūtaḥ}	

Table: Machine internal representation of Analysis

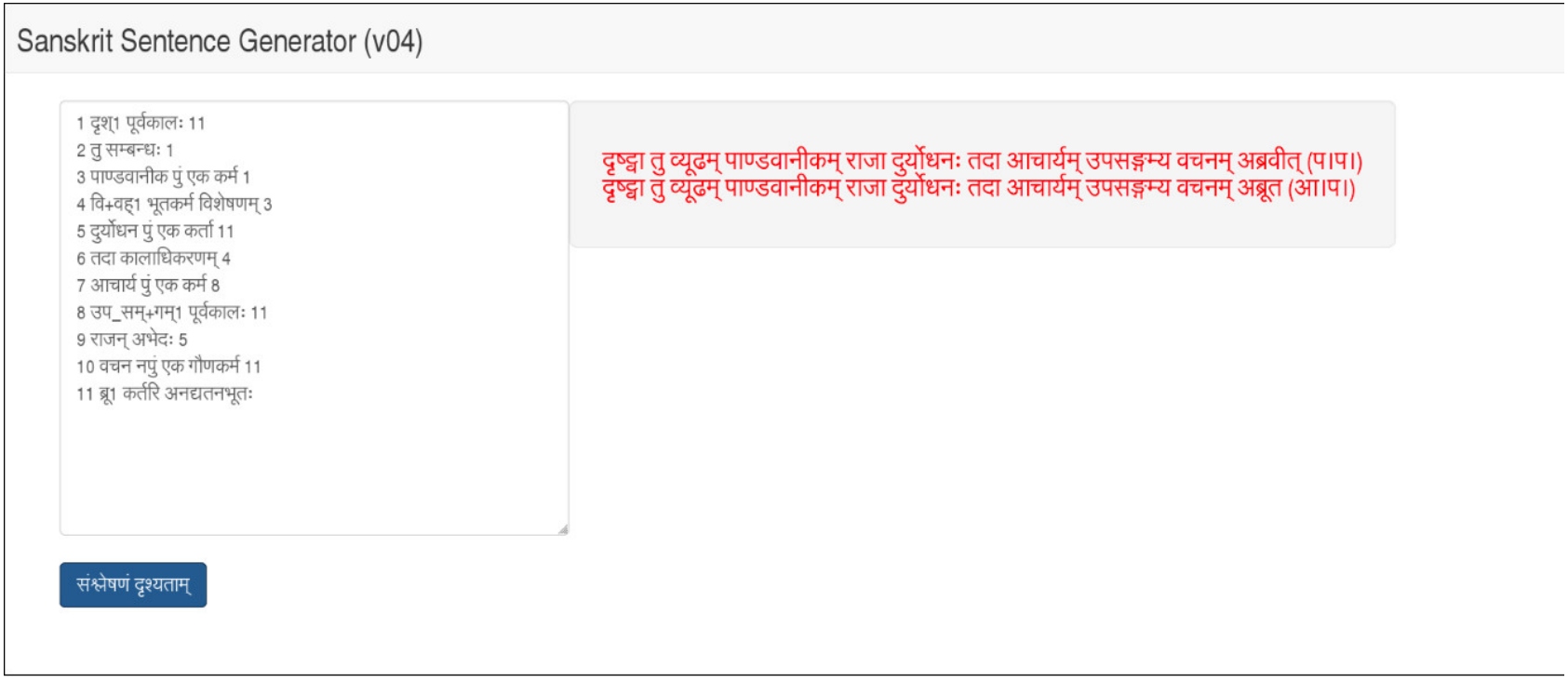


Figure: Generation of BhG 1.2 Shloka from its analysis

A tagset is appropriate for analysis as well as generation, provided

- it is suitable for computer parsing with high accuracy
- it is also easily comprehensible by a non-grammarian

This tagset has two problems from the generation point of view

- Some of the relations are morpho-syntactic
  - upapada-sambandhaḥ
  - ṣaṣṭhī-sambandhaḥ
- Special sūtras by Pāṇini in Kāraka section for certain verbs with / without prefixes are un-natural for a non-grammarian
  - adhiśīṡthāsāṁ karma (1.4.46)
  - krudh-druh-īrṛā-asūyā-arthānāṁ yaṁ prati kopaḥ (1.4.37)

Upapada: The word placed near / an adjoining word
The case marker governed by upapada: upapada-vibhakti

According to the Consortium guidelines, the analysis of the sentence

grāmam paritaḥ vṛkṣāḥ santi

was

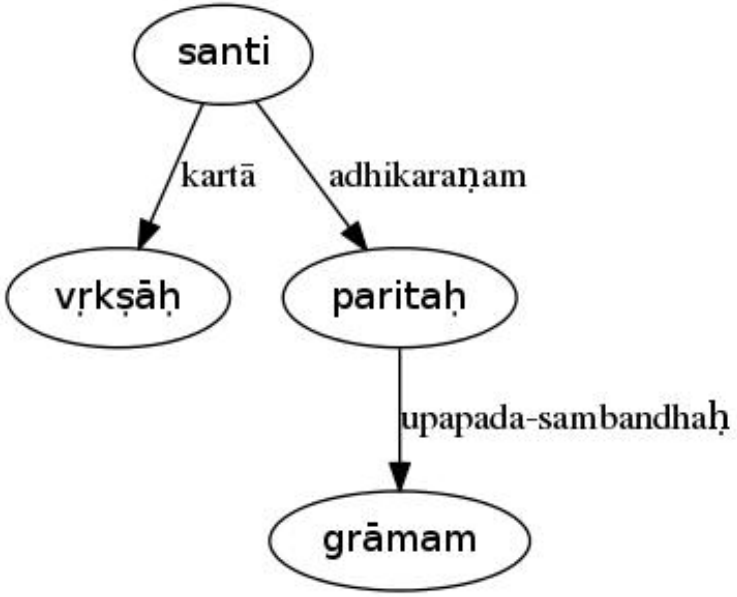


Figure: Morpho-syntactic analysis of (1)

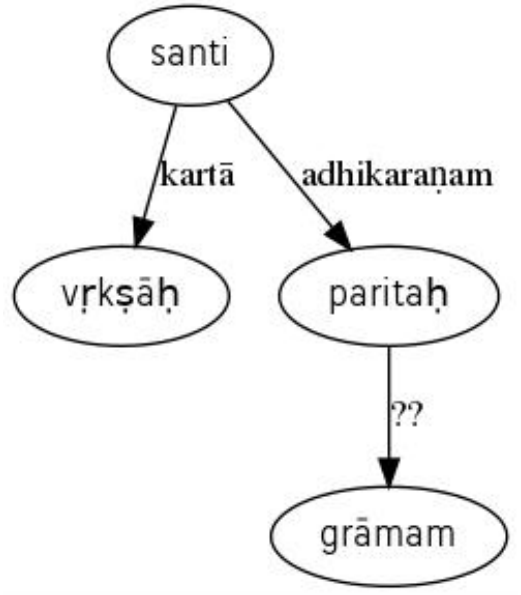
How natural it is to comprehend this analysis?

Requires good knowledge of grammar!

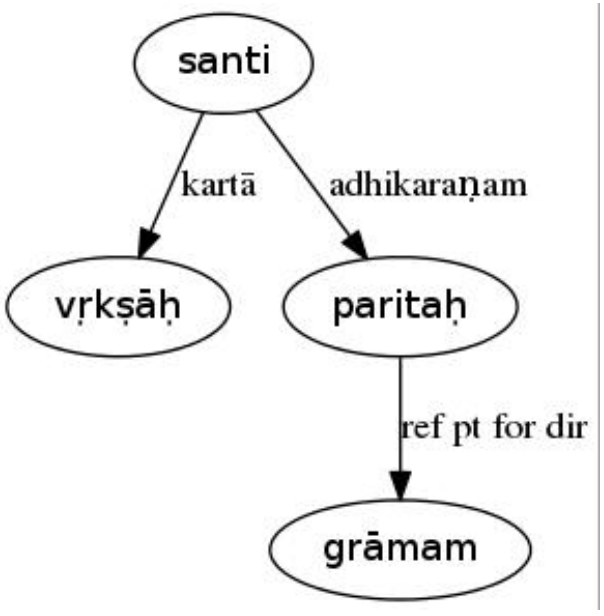
Analysis: not comprehensible

For generation, it is unrealistic to expect a user to provide *upapadasambandhaḥ* as a relation !

grāmam paritaḥ vṛkṣāḥ santi

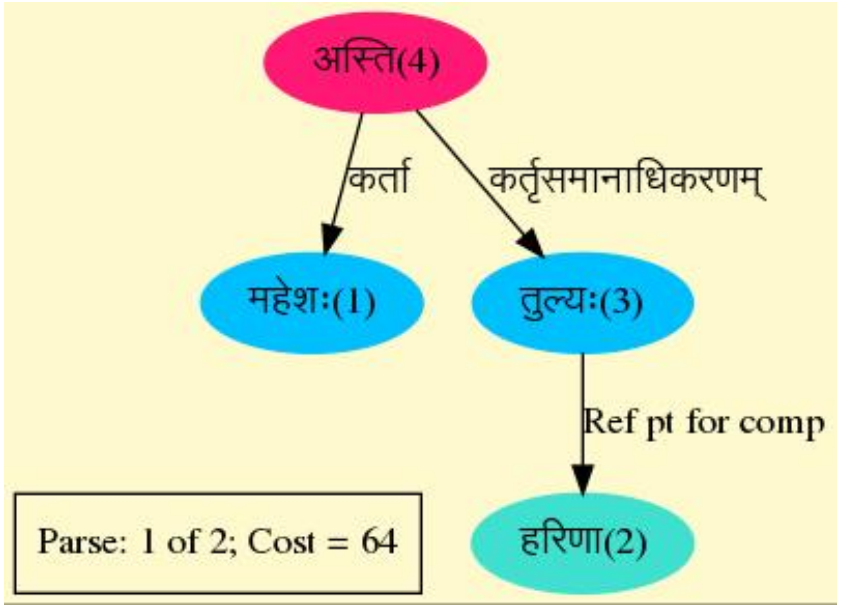


Reference point for direction



grāmam paritaḥ vṛkṣāḥ santi

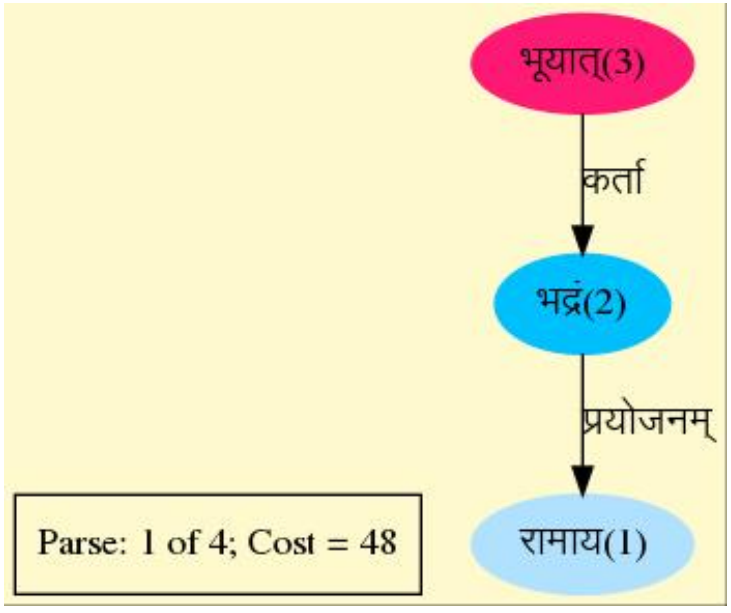
Point of reference for comparison:



maheśaḥ hariṇā tulyaḥ asti.

Prayojanam

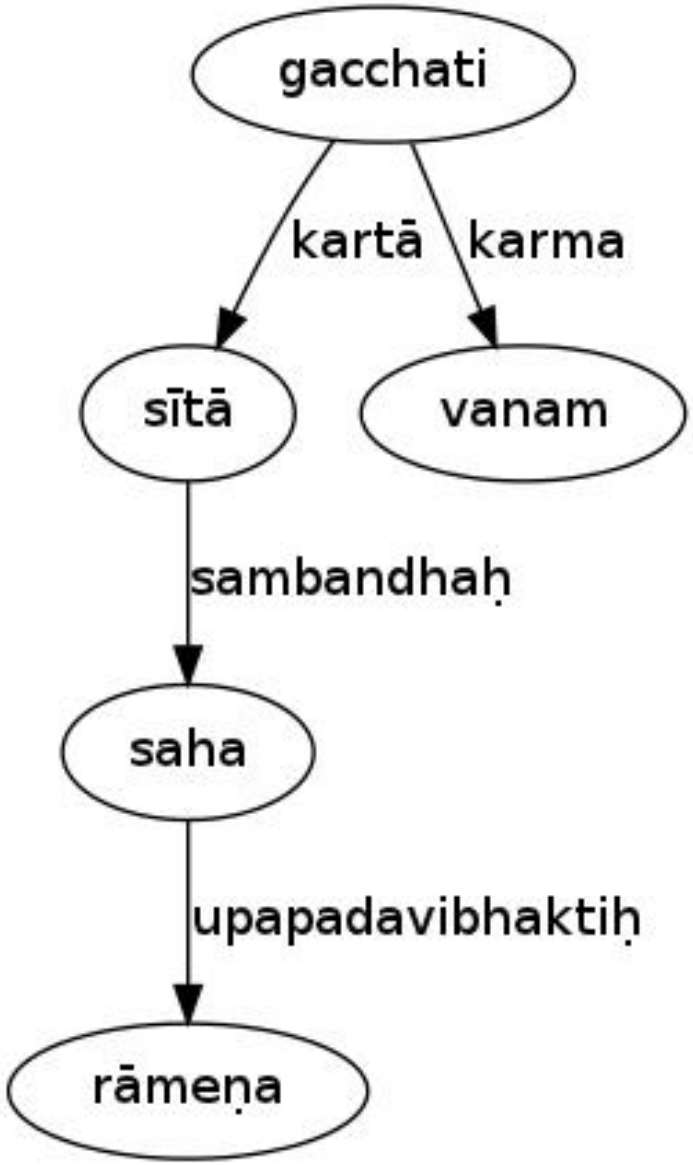
rāmāya bhadraṁ bhūyāt



Association

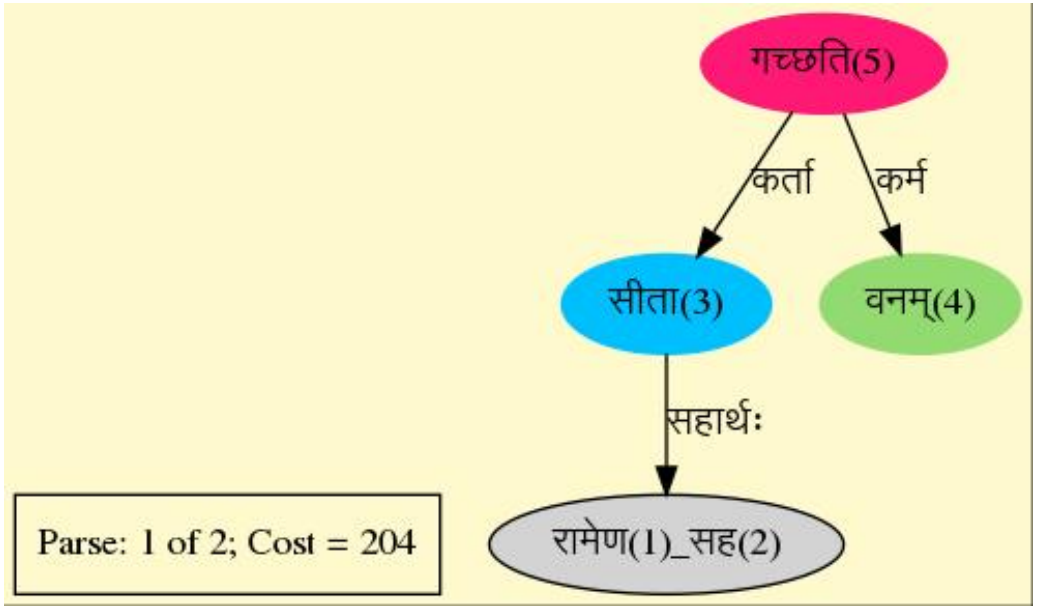
Sītā rāmeṇa saha vanaṁ gacchati

According to the Consortium tagging guidelines



Association

Sītā rāmeṇa saha vanaṁ gacchati



Semantics associated with upapadas

1. Reference point for direction,
   paritaḥ, abhitaḥ, purataḥ, adho'dhaḥ, adhastāt, purastāt, ...
2. Reference point for comparison,
   tulya, sama, samāna, sadṛśa, ...
3. Other relations
   - association,
   - prayojanam,
   - master / owner, ...
4. Karmapravacanīyas: words which make known an action
   - hetu-hetumadbhāva
   - lakṣaṇa (sign / indication)
   - sahita (along with)
   - hīna (inferiority)
   - varjana (exclusion)
   - ādhikya (excess)
   - maryādā (limit)
   - ittambhūtākhyāna
   - pūjā (worship)
   - atikramaṇa (surpassing)
   - īśvara (lord), ...

Conclusion

- The tagset is now purely semantic
  $\Rightarrow$ The relations are language independent

Dhanyavādaḥ