Geeta: Gold Standard Annotated Data, Analysis and its Application

Preeti Shukla
Amba Kulkarni
Devanand Shukl

Department of Sanskrit Studies,
University of Hyderabad,
Hyderabad

20/12/2013

Outline

Gold standard data

When a corpus is manually annotated at various levels, such as Tree bank annotation, Discourse level annotation, etc., then it serves as a gold standard data for evaluation and comparison of various tools. This has the following advantages –

Advantages

1. One can use the gold standard annotated data as an input for the evaluation of various modules. This avoids the cascading effects and one can measure the absolute performance of various tools.
2. To use data-driven models to tune the machine for better performance in a chosen domain.
   For eg., Penn Tree bank

Advantages

1. One can use the gold standard annotated data as an input for the evaluation of various modules. This avoids the cascading effects and one can measure the absolute performance of various tools.
2. To use data-driven models to tune the machine for better performance in a chosen domain.
   For eg., Penn Tree bank

Śrīmad Bhagvad Gītā

We chose Śrīmad Bhagvad Gītā (BhG for short) consisting of 18 chapters and 700 verses (ślokas) for developing a gold standard data for Sanskrit text mainly because:

- It is an important text which summarizes the Upanishadic teachings and is commented upon and interpreted by various schools of Indian philosophies.
  Thus annotators in doubt can always refer to these commentaries for correct annotation.

- This scripture being coherent and complete in itself, can be used for higher level analysis such as discourse analysis, topic identification, anaphora resolution, and so on.

- Further, this also being part of the Mahābhārata (section 25 to 42 of the Bhiṣma parva), later on if necessary, it can be used as an initial training data for boot-strapping for automatic annotation of complete critical edition of Mahābhārata (with around hundred thousand verses).

Sanskrit requires the following levels of tagging viz.,

1. Annotation of Sandhi
2. Tagging of Compounds
3. Morphological Analysis
4. Tagging of Sentential relations
5. Marking the Prose order

Sanskrit requires the following levels of tagging viz.,

1. Annotation of Sandhi
2. Tagging of Compounds
3. Morphological Analysis
4. Tagging of Sentential relations
5. Marking the Prose order

Sanskrit requires the following levels of tagging viz.,

1. Annotation of Sandhi
2. Tagging of Compounds
3. Morphological Analysis
4. Tagging of Sentential relations
5. Marking the Prose order

Sanskrit requires the following levels of tagging viz.,

1. Annotation of Sandhi
2. Tagging of Compounds
3. Morphological Analysis
4. Tagging of Sentential relations
5. Marking the Prose order

Sanskrit requires the following levels of tagging viz.,

1. Annotation of Sandhi
2. Tagging of Compounds
3. Morphological Analysis
4. Tagging of Sentential relations
5. Marking the Prose order

Annotation of Sandhi

The first level of tagging needed for Sanskrit is the marking of word boundaries undoing the sandhi (euphonic changes). We split two types of sandhis –

1. sandhi between two words which is indicated by a '+' sign
2. sandhi between the components of a compound is indicated by a '-' sign

Annotation of Sandhi

The first level of tagging needed for Sanskrit is the marking of word boundaries undoing the sandhi (euphonic changes). We split two types of sandhis –

1. sandhi between two words which is indicated by a '+' sign
2. sandhi between the components of a compound is indicated by a '-' sign

Example

BhG 2.40 –

nehābhikramanāśo'asti pratyavāyo na vidyate|
svalpamapyasya dharmasya trāyate mahato bhayāt||

Eng: In this path there is no loss of effort, nor is there any adverse result. Even a little practice of this discipline protects one from great fear (of birth and death).

Verse tokenized as

na+iha+abhikrama-nāśaḥ+asti pratyavāyaḥ+na *vidyate*|
svalpam+api+asya *dharmasya trāyate* ***mahataḥ+bhayāt***||

Tagging of Compounds

Sanskrit compounds are broadly classified as follows:

1. endo-centric / adverbial (tatpuruṣaḥ / avyayībhāvaḥ)
2. exo-centric (bahuvrīhiḥ)
3. copulative (karmadhārayaḥ)
4. conjunctive (dvandvaḥ)

These compounds with an exception of dvandva (conjunctive) and bahuvrīhi (exocentric) are binary.

Tagging of Compounds

Sanskrit compounds are broadly classified as follows:

1. endo-centric / adverbial (tatpuruṣaḥ / avyayībhāvaḥ)
2. exo-centric (bahuvrīhiḥ)
3. copulative (karmadhārayaḥ)
4. conjunctive (dvandvaḥ)

These compounds with an exception of dvandva (conjunctive) and bahuvrīhi (exocentric) are binary.

Tagging of Compounds

Sanskrit compounds are broadly classified as follows:

1. endo-centric / adverbial (tatpuruṣaḥ / avyayībhāvaḥ)
2. exo-centric (bahuvrīhiḥ)
3. copulative (karmadhārayaḥ)
4. conjunctive (dvandvaḥ)

These compounds with an exception of dvandva (conjunctive) and bahuvrīhi (exocentric) are binary.

Tagging of Compounds

Sanskrit compounds are broadly classified as follows:

1. endo-centric / adverbial (tatpuruṣaḥ / avyayībhāvaḥ)
2. exo-centric (bahuvrīhiḥ)
3. copulative (karmadhārayaḥ)
4. conjunctive (dvandvaḥ)

These compounds with an exception of dvandva (conjunctive) and bahuvrīhi (exocentric) are binary.

The meaning of a compound is decided by the way the components are combined together.
A compound with 3 components a-b-c may be combined in two different ways viz.,

- <<a-b>-c>
- <a-<b-c>>

Example

The compound word dehāntaraprāptiḥ in BhG 2.13 is combined as
<<**a-b**>-**c**>
anyaḥ dehaḥ – dehāntaram
Gloss: transference_of_the_body
dehāntarasya prāptiḥ – dehāntaraprāptiḥ
Gloss: attainment_of_transference_of_the_body

Example

The compound word dehāntaraprāptiḥ in BhG 2.13 is combined as <<**a-b**>-**c**>

anyaḥ dehaḥ – dehāntaram

Gloss: transference_of_the_body

dehāntarasya prāptiḥ – dehāntaraprāptiḥ

Gloss: attainment_of_transference_of_the_body

Various stages

Kumar et al., (2010) describe various stages involved in the analysis of a compound which also form the natural modules of a compound processor –

1. Segmentation (samāsapadacchedadh)
2. Constituency Parsing (samāsapadānvayaḥ)
3. Compound Type Identification (samastapadaparicāyakaḥ)
4. Paraphrasing (vigraha-vākyam)

A hierarchical tagset of 55 tags has been designed to tag the Sanskrit compounds.

Various stages

Kumar et al., (2010) describe various stages involved in the analysis of a compound which also form the natural modules of a compound processor –

1. Segmentation (samāsapadacchedadh)
2. Constituency Parsing (samāsapadānvayaḥ)
3. Compound Type Identification (samastapadaparicāyakaḥ)
4. Paraphrasing (vigraha-vākyam)

A hierarchical tagset of 55 tags has been designed to tag the Sanskrit compounds.

Various stages

Kumar et al., (2010) describe various stages involved in the analysis of a compound which also form the natural modules of a compound processor –

1. Segmentation (samāsapadacchedadh)
2. Constituency Parsing (samāsapadānvayaḥ)
3. Compound Type Identification (samastapadaparicāyakaḥ)
4. Paraphrasing (vigraha-vākyam)

A hierarchical tagset of 55 tags has been designed to tag the Sanskrit compounds.

Various stages

Kumar et al., (2010) describe various stages involved in the analysis of a compound which also form the natural modules of a compound processor –

1. Segmentation (samāsapadacchedadh)
2. Constituency Parsing (samāsapadānvayaḥ)
3. Compound Type Identification (samastapadaparicāyakaḥ)
4. Paraphrasing (vigraha-vākyam)

A hierarchical tagset of 55 tags has been designed to tag the Sanskrit compounds.

Morphological Analysis

At this stage tagging of both the inflectional as well as derivational information is needed. e.g.,
The morph of the verb ‘asti’ (to be) is

V

as2{active;laṭ;1st;sg;parasmaipadī;asaz;adādiḥ}

Tagging of Sentential relations

In order to interpret the meaning of a sentence, various relations among words are necessary.
Of around 90 relations classified by Ramakrishnamacharyulu (2009), only around 35 relations based on syntactico-semantic information available in a sentence are considered for automatic tagging.

Marking the Prose order

For understanding a Sanskrit text in verse style, two different methods have been followed in Indian education system viz.,

- **Daṇḍānvaya** (also known as anvayamukhī)
  The teacher arranges all the words in prose order for easy understanding of a verse.
  This approach assumes that if a user is given the 'default prose order' of the sentence, he 'understands' its meanings.

Example

BhG 4.8 –

paritrāṇāya sādhūnāṃ vināshāya ca duṣkṛtām|
dharmasaṃsthāpanārthāya sambhavāmi yuge yuge||

Anvaya: *aham sādhūnāṃ paritrāṇāya duṣkṛtām vināshāya dharmasaṃsthāpanārthāya ca yuge yuge sambhavāmi*

Eng: I appear from time to time for protecting the good, for transforming the evil-minded, and for establishing world order (Dharma).

- **Khaṇḍānvaya** (also known as kathambhūtinī)

  The teacher gives the basic skeleton of a sentence and fills in other details by asking questions which are centered around the heads seeking their various modifiers.

  This approach is close to parsing a sentence showing various dependency relations.

Canonical Form

The default word order or the 'canonical form' is governed roughly by the following verse:

samāsacakram kā.verse 10

viśeṣaṇam puraskṛtya viśeṣyam tad-lakṣaṇam|
kartṛ-karma-kriyā-yuktam etad anvaya-lakṣaṇam||

gloss: Starting with the adjectives, targeting the headword, in the order of kartṛ-karma-kriyā (subject-object-verb) gives an anvaya.

Example – BhG 2.40

<।>
<seg type="pāda">
<euphonic-word no="1"> nehābhikramanāśo'sti
<word no="1" prose word order_no="3"> na
<mo_anal> na{ind} </mo_anal>
<syntactic_rel> mod 4 </syntactic_rel>
</word>
<word no="2" prose word order_no = "1"> iha
<mo_anal> iha{ind} </mo_anal>
<syntactic_rel> loc 4 </syntactic_rel>
</word>

<word no="3" prose word order_no="2"> abhikramanāśaḥ
<compound label="T6">
<component no="1"> abhikrama </component>
<component no="2"> nāśaḥ
<mo_anal> nāśa{masc}{nom;sg} </mo_anal>
</component>
</compound>
<syntactic_rel> subj 4 </syntactic_rel>
</word>
<word no="4" prose word order_no="4" > asti
<mo_anal>
as2{active;laṭ;1st;sg;parasmaipadī;asaz;adādiḥ} </mo_anal>
</word>
</euphonic-word>
</seg>
</l>

The process for semi-automatic tagging of BhG is as follows:

- The verse form is converted into prose form.

- Initially the sandhi and compound in the verse are segmented manually, following the guidelines developed by the SHMT consortium [1]. Then each compound is tagged for its type, along with the complete constituency mark-up.

[1] This is the Consortium of 7 institutes, for 'Development of Sanskrit-Hindi Machine Translation System (sampark)' funded by DIT, Govt. of India

- The segmented words are run in the anusāraka interface [2] for obtaining the multiple morph analysis. The output generated as an xml file, is then manually pruned for choosing the correct morph analysis in the context.

[2] http://sanskrit.uohyd.ernet.in/scl

- The synactico-semantic relations are tagged manually, following the guidelines develped by the SHMT consortium [3].

[3] http://sanskrit.uohyd.ernet.in/scl/Corpus/TaggingGuidelines/kaaraka-tagging-guidelines

- Hindi and English glosses for each word are given manually. For this we followed *Geeta Press* [4].

[4] Śrīmad Bhagvad Gītā, Geeta Press, Gorakhpur, India, reprint 2007

Compound Distribution

Compound-type	Freq
Endocentric	994
Exocentric	390
Copulative	163
Conjunctive	144

◂ Refer DSP

Morphological statistics:

multiple morph count	words	multiple morph count	words
1	5280	8	22
2	1570	9	28
3	1082	10	15
4	349	11	10
5	292	12	7
6	99	13	3
7	121	14	6
		Total words	8884
		Average	1.90

Case Statistics in Geeta

case	singular	dual	plural
nom.	2463	31	613
acc.	1349	20	251
instr.	266	0	94
dat.	57	0	5
abl.	116	0	12
gen.	335	24	179
loc.	273	3	92
voc.	251	1	0

Case-Number distribution in context

case	singular	dual	plural
nom.	864	29	501
acc.	733	20	103
instr.	78	0	48
dat.	29	0	5
abl.	53	0	12
gen.	222	24	98
loc.	155	3	48
voc.	80	1	0
total	2014	74	795

◂ Refer DSP

syntactico-semantic relations in Geeta

relation	freq	relation	freq
viśeṣaṇa (adjective)	1277	kartā (subject)	1256
conjunctive	1155	karma (object)	924
predicative adj	401	adhikaraṇa (locative)	358
ṣaṣṭhi (genitive)	357	negation	242
emphatic	275	sambodhana (vocative)	237
precedence	194	adverb	194
karaṇa (instrument)	130	karma-sāmānidhikaraṇa	113
hetu (causal)	96	co-relative sarvanāma (pronouns)	82
apādāna (source)	75	vākya-karma	64
prayojana (purpose)	52	simultaneity	50
sampradāna (dative)	15	disjunction	10

Tense-Mood distribution in Geeta

Sanskrit has 10 lakāras which represent Tense-Modality.

lakāra	freq
laṭ (Present)	355
loṭ (Imperative)	72
lṛṭ (Second future)	42
vidhiliṅ (Potential)	40
laṅ (Imperfect)	26
liṭ (Perfect)	22
luṭ (First future)	16
luṅ (Aorist)	9
āśīrliṅ (Optative)	6
lṛṅ (Conditional)	1

◂ Refer DSP

Utility

There are three important usages of this gold data viz.,

1. for NLP applications
2. as linguistic inputs for developing a domain specific primer for learning BhG
3. with the suitable interface, a self-learning / reading tool for BhG.

Utility

There are three important usages of this gold data viz.,

1. for NLP applications
2. as linguistic inputs for developing a domain specific primer for learning BhG
3. with the suitable interface, a self-learning / reading tool for BhG.

Utility

There are three important usages of this gold data viz.,

1. for NLP applications
2. as linguistic inputs for developing a domain specific primer for learning BhG
3. with the suitable interface, a self-learning / reading tool for BhG.

NLP applications

- serves as a gold standard for evaluation of various Sanskrit tools.
- for developers of NLP tools.
- in prioritizing the solutions in the case of ambiguities.

NLP applications

- serves as a gold standard for evaluation of various Sanskrit tools.
- for developers of NLP tools.
- in prioritizing the solutions in the case of ambiguities.

NLP applications

- serves as a gold standard for evaluation of various Sanskrit tools.
- for developers of NLP tools.
- in prioritizing the solutions in the case of ambiguities.

NLP applications

- serves as a gold standard for evaluation of various Sanskrit tools.
- for developers of NLP tools.
- in prioritizing the solutions in the case of ambiguities.

Domain Specific Primer

- The quantitative analysis can help a teacher to decide which aspect of Sanskrit Grammar is more relevant for the study of BhG as follows:
  1. postponing the teaching of dual forms to a later stage.
  2. deciding how many and which paradigms of noun declension to concentrate on first.
  3. using the analysed data of Tense-Modality to decide which lakāras to teach and which conjugation classes to concentrate on first.
  4. deciding on which type of compound to teach first based on the compound distribution table.

DSP ▸ Go to Case ▸ Go to lakara ▸ Go to CD

Domain Specific Primer

- The quantitative analysis can help a teacher to decide which aspect of Sanskrit Grammar is more relevant for the study of BhG as follows:
  1. postponing the teaching of dual forms to a later stage.
  2. deciding how many and which paradigms of noun declension to concentrate on first.
  3. using the analysed data of Tense-Modality to decide which lakāras to teach and which conjugation classes to concentrate on first.
  4. deciding on which type of compound to teach first based on the compound distribution table.

DSP ▸ Go to Case ▸ Go to lakara ▸ Go to CD

Self Learning cum Reading Tool

With the help of a suitable interface such as that of anusāraka, an interested reader can have complete analysis of BhG at various levels. The interface provides –

- User controlled access to various levels of analysis (Figs. 1,2,3). The graphs showing the constituency information and the kāraka relations are generated automatically from the manually tagged data.

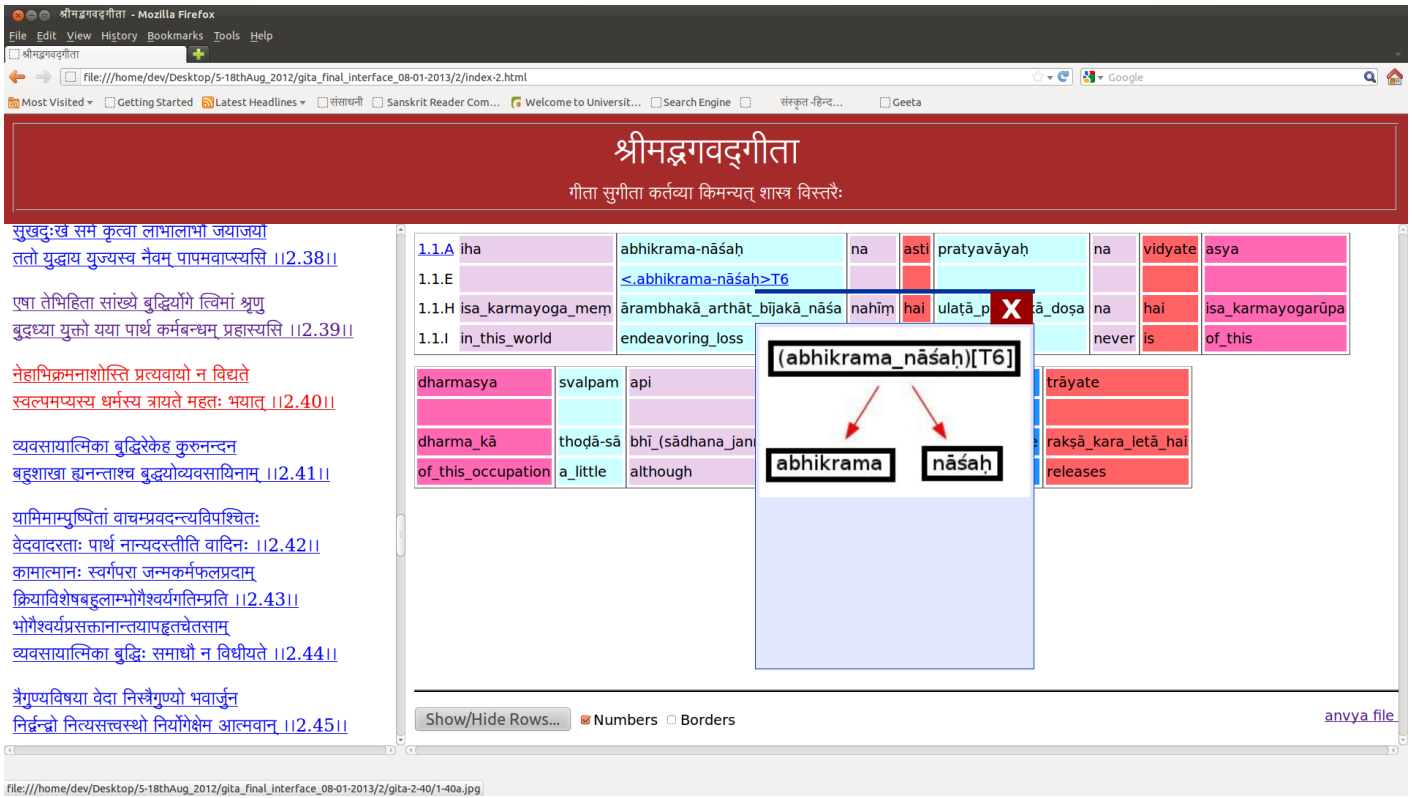

Figure: 1. compound analysis of BhG 2.40

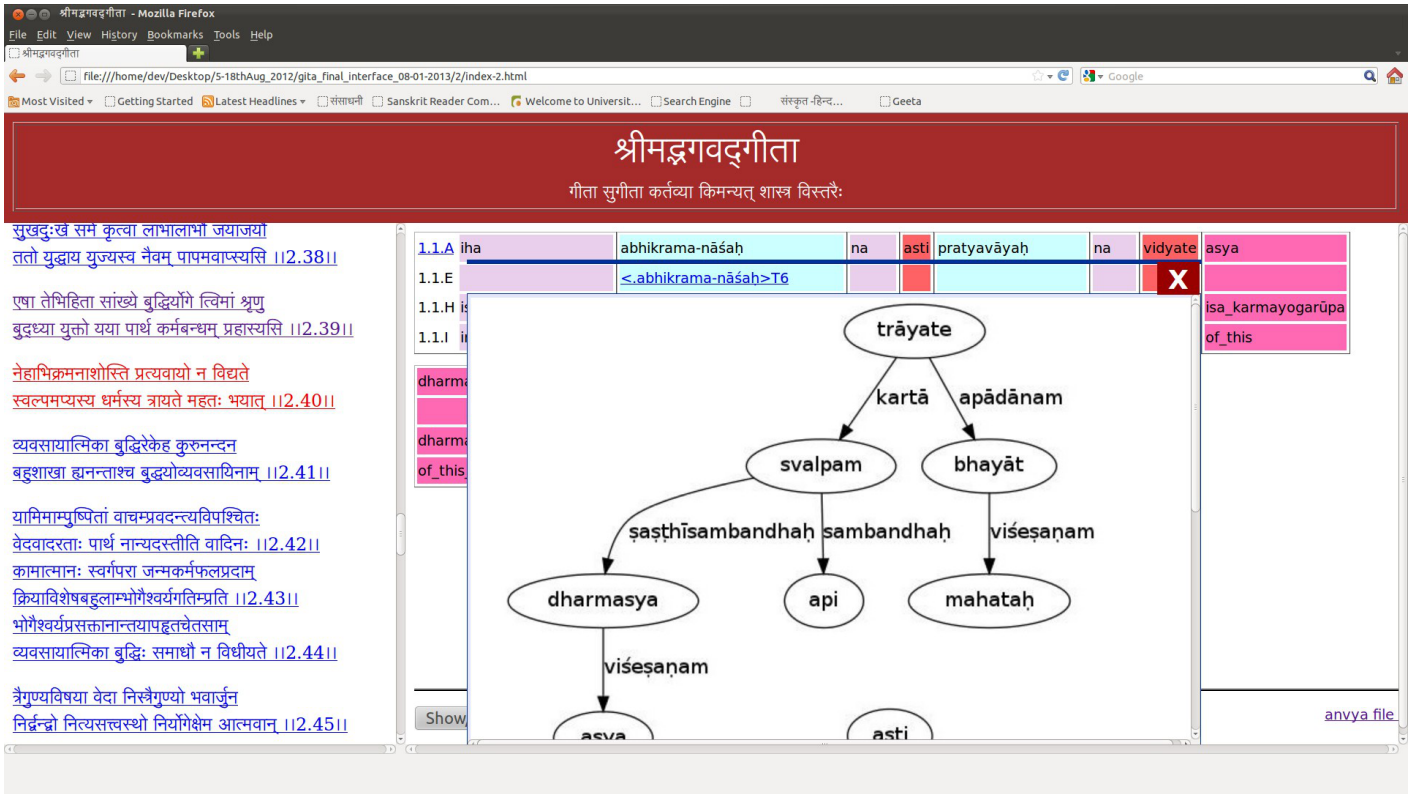

Figure: 2. kāraka analysis of BhG 2.40

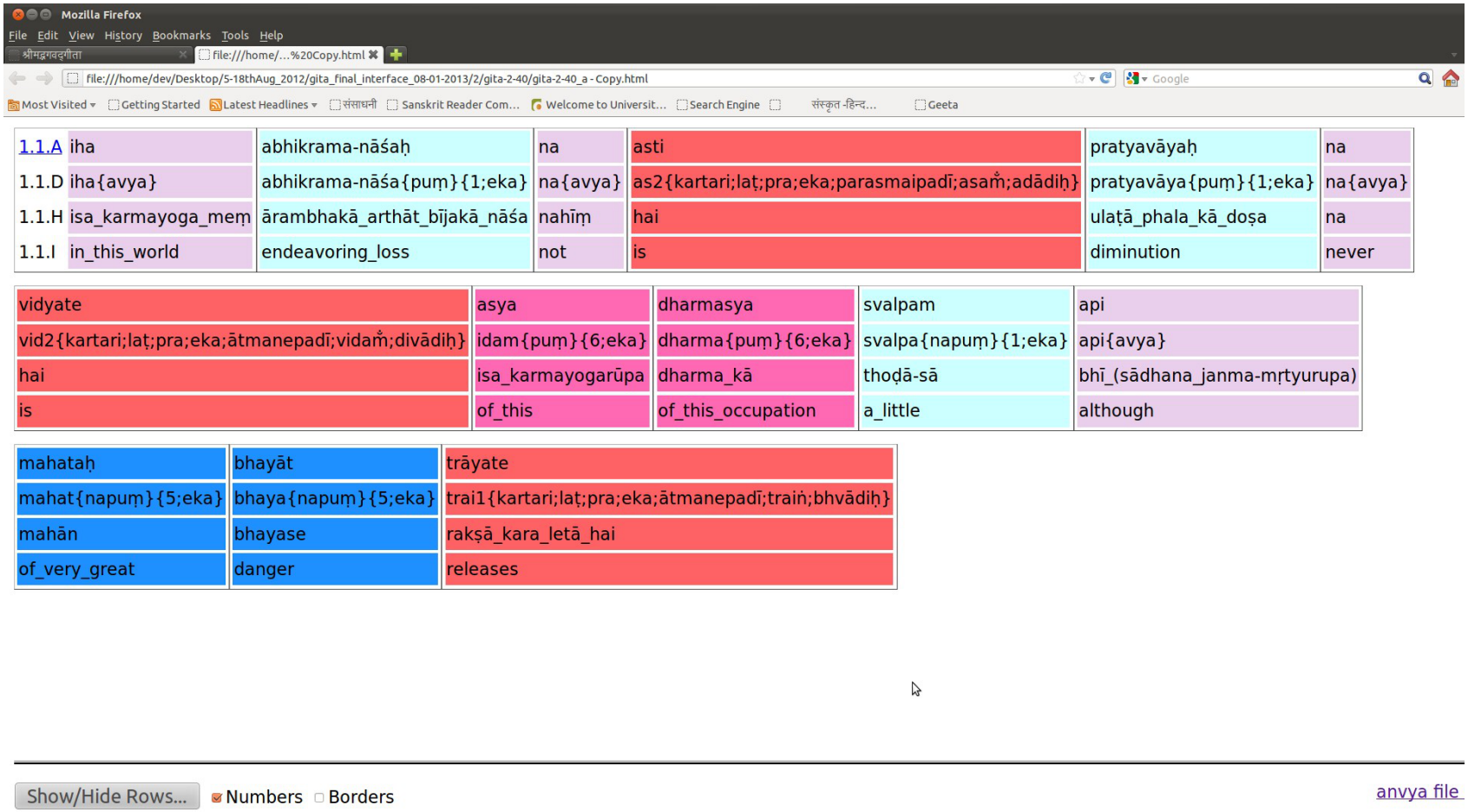

Figure: 3. morph analysis of BhG 2.40

- Link to various dictionaries for meanings of the head words.
- Graphical display of phrase structure analysis of compounds (Fig 4).

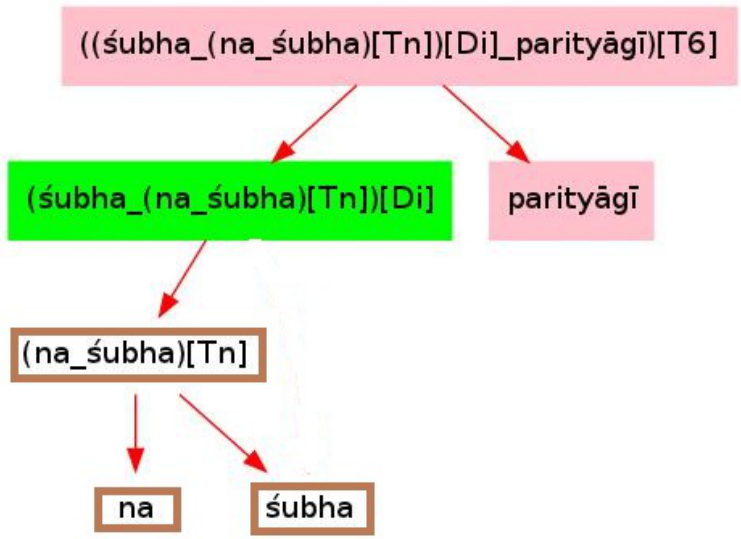

Figure: 4. Compound analysis of a word from BhG 12.17

- Graphical display of sentential analysis (Fig 5).

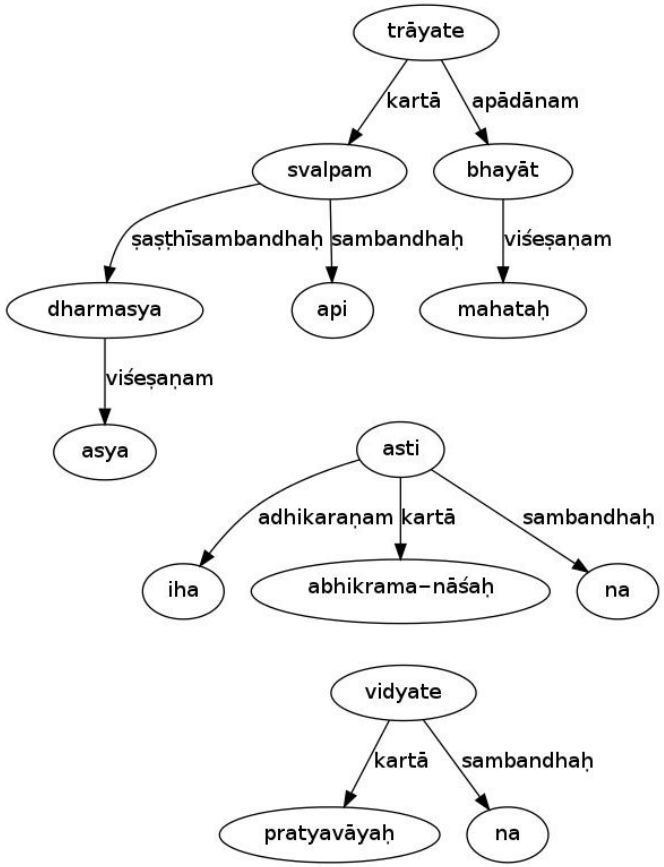

Figure: 5. Dependency graph of BhG 2.40

This provides the user a digitized learning and understanding environment and forms a basis for the theoretical linguists and grammarians to test their theories as well.

धन्यवादः!!!