Use of Features for Accentuation of *ghañanta* Words

Samir Janardan Sohoni
IIT Bombay
Powai, Mumbai 400076
India
sohoni@hotmail.com

Malhar A. Kulkarni
IIT Bombay
Powai, Mumbai 400076
India.
malharku@gmail.com

Abstract

Sanskrit is an accented language. The accent is very prominent in Vedic Sanskrit but classical Sanskrit does not make use of it. Accent is a disambiguation device which can help to understand the correct sense of a word.

Words in Sanskrit tend to fall in certain grammatical categories like *ghañanta*s, vocatives etc. Words in a category tend to have an accent governed by either a general or a special rule of accentuation. Pāṇini's special (*apavāda*) rules of accentuation for a specific category of words can be studied along with the general (*utsarga*) rules of accentuation for the same class. The resulting body of rules states all the conditions necessary for accentuation of words from that class. The conditions reveal a set of features which can be used for determining the correct accent. If the features of accentuation for a class of words is understood, computers can be trained to accentuate unaccented words from that class. This paper discusses some features of *ghañanta* words and how the features can be used to restore accent on unaccented *ghañanta* words.

1 Introduction

Broadly, Sanskrit has two streams – Vedic and Classical. The latter is known as *bhāṣā*, the commonly spoken language which is the main focus of Pāṇini's grammar. The former is indicated in Pāṇini's grammar by terms like *mantre* and *chandasi*. Accentual contrasts are one distinctive feature of Vedic. During the time Pāṇini developed his Sanskrit grammar, accentuation must have been a feature of the language. Pāṇini has made provision for accents in the morphological machinery of Sanskrit – the roots (*dhātu*s) and suffixes (*pratyaya*s). The Aṣṭādhyāyī lays out the mechanisms for production of Vedic forms as well as those in bhāṣā using the same morphological framework. Accent was destined to vanish from written form of classical texts – none of the Sanskrit classics use accent. Classics would not have been able to use certain *alaṅkāra*s if accentual contrasts were in force[1].

Accent plays an important role in ascertaining the senses of words. For example, the word *jyeṣṭháḥ*[2] means *elder* when pronounced with an acute final syllable. If pronounced as *jyéṣṭhaḥ*[3], which has an acute initial syllable, it means *praiseworthy*. Both forms are created by adding the superlative suffix *iṣṭhan*. Another example is the base *arya* which means *honourable*, *lord* or someone earning a living by doing business. If the word is pronounced as *aryá*[4] it strictly means a lord. If pronounced as *árya* it can take other meanings.

Although classical Sanskrit lacks tonal contrasts, marking plain text with correct accents can lower ambiguity and thereby contribute to high-quality translations of classical Sanskrit texts. Marking plain text with correct accents can help to reduce ambiguity by narrowing down senses of words. Low ambiguity can result in high-quality translations.

Pāṇinian grammar is generative – initial

[1] Devasthali (1967) pp 1 note 1 – *kāvyamārge svaro na gamyate*

[2] The derivation begins as *vṛddha* + *iṣṭhan*. See Aṣṭādhyāyī rule *jya ca* (A. 5.3.61). The accent is by *phiṭhsūtra*s, see *jyeṣṭhakaniṣṭhayoḥ vayasi* (P. 1.23).

[3] The derivation begins as *praśasya* + *iṣṭhan*. See Aṣṭādhyāyī rule *vṛddhasya ca* (A. 5.3.62). The accent is by *ñnityādirnityam* (A. 6.1.197).

[4] See *aryasya svāmyākhyā cet* (P. 1.18) and *aryaḥ svāmivaiśyayoḥ* (A. 3.1.103).

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 229–238, Varanasi, India. December 2016.

conditions coupled with the rules from the Aṣṭādhyāyī lead to the final accented form. This might be the only way Pāṇinian grammar is designed to work. How, then, can accent be marked on unaccented words, such as those found in classical texts, using the Pāṇinian process? It seems that words have features due to which they are accentuated a certain way. If these features are learned and superimposed on unaccented words, they can be given the correct accent.

2 Notion of Compositionality

The meaning and accent of a sentence are composed of the meaning and the accent of words. Similarly, the meaning and accent of a word are composed of the meaning and the accent of morphological elements like roots and terminations. Kulkarni et al. (2015) have shown that sentences are based on the notion of compositionality.

An example of this notion of compositionality is shown in figure 1. Unlike sentential meaning, sentential accent is not cumulative. Sentential accent is not simply a concatenation of accented words. Accented words such as *gopāláḥ* (an agent), *grāmáṃ* (an object) and *gácchati* (a verb) can be combined to form a sentence which means 'Gopāla goes to town.' However, the concatenated form *gopāláḥ grāmáṃ gácchati* is not a correctly accented sentence. The correct rendering[5] is *gopāláḥ grāmáṃ gacchati* because the rule *tiṅ atiṅaḥ* (A. 8.1.28) makes all syllables of a finite verb low-pitched when it follows something other than a finite verb. Just like *tiṅ atiṅaḥ* provides special syntactic conditions for accentuation of finite verbs, other rules lay down semantic, syntactic and lexical conditions for accentuation of many classes of words.

The accent on finite verbs is subject to several sentence-level syntactic considerations (See sections 8.1 and 8.2 of the Aṣṭādhyāyī). The rule *tiṅ atiṅaḥ* (A. 8.1.28) happens to be one rule among many which regulates verbal accent. This rule looks at the internal structure of adjacent words to decide if a verb follows a non-verb. Some rules governing accent on compounds also look at

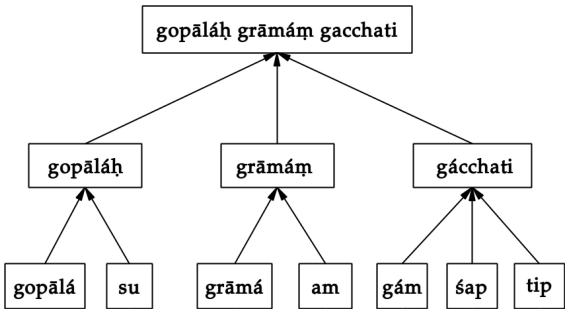

Figure 1: An example of how sentential accent develops

internal structure of words. For example, the compound *mudga-cū̃rṇam* (paraphrased as *mudgasya cūrṇam*) is accentuated by the rule *cūrṇādīnyaprāṇiṣaṣṭhyāḥ* (A. 6.2.134). The genetive suffix *sya* is elided in the process of compound formation. The rule A. 6.2.134 needs to know if the first word of the compound had a genetive suffix. Thus, the internal structure of words is important.

3 The Deep Structure of Words

The surface form of a word is the sequence of phonemes, which are represented by graphic letters. The word *rājā* is made up of the string of letters $r + \bar{a} + j + \bar{a}$. The surface form is the final form used in language.

For accentuation, meta information used by Pāṇinian rules is required. The form of the word which contains this meta information can be called the deep structure of the word. The meta information annotated in the deep structure is of two kinds. One kind represents the *saṃjñā*s and the second kind represents the transformations of morphological elements.

3.1 The *saṃjñā* Labels

As a word is derived using Pāṇinian rules, several *saṃjñā*s are applied to various parts of the word. These are Pāṇinian labels used in grammar to identify morphologicial elements so that operations can be applied to those elements. These labels, such as *dhātu* (root), *pratyaya* (suffix), etc. are pieces of meta information that belong to the deep structure of the word. Any keyword, which is important for a rule of accentuation, can be annotated onto the corresponding substring of the word in the deep structure.

[5]Roman transliteration uses grave (̀) for circumflex (*svarita*) accent and acute (́) for the acute (*udātta*).

3.2 The Transformations

In Pāṇinian grammar, the derivation of a word begins from an initial set of morphological affixes. The morphological elements undergo a transformation to become bits and pieces of the final form of the word. Often, a rule of accentuation specifies accent on an affix and that accent is honoured even if the affix is transformed later in the derivation. For instance, in the third person singular *lṛṭ* (second future) form *bhaviṣyáti* (will become), the typical suffix *sya* is introduced by the rule *syatāsī lṛluṭoḥ* (A. 3.1.33) and it has an acute initial syllable due the rule *ādyudāttaḥ ca* (A. 3.1.3). Subsequently, *sya* transforms into *ṣya* due to the rule *ādeśapratyayayoḥ* (A. 8.3.59), but retains the acute of the original suffix *sya*.

In general there are six different types of transformations in Pāṇinian grammar as shown in Figure 2. The initiation operation (fig. (a)) refers to the first use (*upadeśa*) of any kind of phonetic entity such as a root or suffix. Carrying forward a phonetic entity without any modification whatsoever is retention (fig. (b)). In retention the label *upadeśa* is not applicable. In modification (fig. (c)), several substitutions (*ādeśa*s) can modify phonetic units as the derivation progresses. Deletion (fig. (d)) refers to Pāṇinian operations that elide phonetic units due to *lopa*, *luk*, *ślu* or *lup*. In disintegration (fig. (e)) *āgama*s (inserted elements) split the original phonetic unit into multiple units. Combination consists of combining multiple phonetic units into a single one due to operations such as *ekādeśa*.

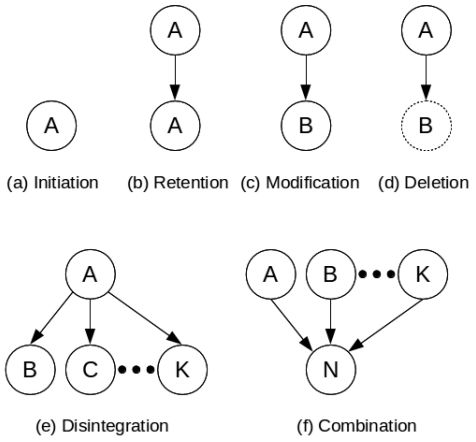

Figure 2: Types of transformations

3.3 Notation of Deep Structure

The transformations are shown using the left arrow. The keywords are shown in square brackets and apply to the immediate phonetic unit on the left. If such a phonetic unit is an aggregate having its own internal deep structure, parentheses are used around it.

As a matter of convention for easy reading, the Devanāgarī script is mixed with Latin script in the annotations of the deep structure. The latter is used only to mark the *saṃjñā*s.

The deep structure of the word *bhaviṣyáti* is shown below.

```
(
  (भ् (अव् <- औ <- ऊ) [udātta])[dhātu]
  ( इ ट्[lopa, it]
    ((ष् <- स्)
    य्
    अ [udātta, active])
  ) [ārdhadhātuka, pratyaya]
  (ति प्[lopa, it]) [pratyaya]
)
```

4 Words Ending in *sup* Suffixes

A sentence is made of words. The string of words is not usually punctuated by spaces in between. Old manuscripts show sentences written with no spaces between words. In Sanskrit there are very few punctuation marks. Perhaps the *daṇda* and the double *daṇda*, which appear at the end of a half and full verse, are the most commonly seen punctuations. The *daṇda*s are ubiquitous because there are no morphological markers for end of the sentence or quarter of a verse.

The rich inflectional morphology of Sanskrit supplies explicit (but non-unique) nominal termintions in seven cases and three numbers per case. The rule *svaujasamauṭ...* (A. 4.1.2) lists these suffixes in their raw form. Together the collection is known as *sup* suffixes. A nominal base can be used only after it is inflected. Such an inflection is called a word (*pada*)[6]. However, due to some mechanisms[7] the *sup* suffixes may vanish blurring the word boundary. However, the status of being a word is retained in spite of this by virtue of rule *pratyayalope pratyayalakṣaṇam* (A. 1.1.62).

[6] Whatever ends in *sup* or *tiṅ* suffixes is a word, cf. *suptighantaṃ padam* (A. 1.4.14)

[7] The rule *halṅyābbhyoḥ dīrghāt sutisyapṛktaṃ hal* (A. 6.1.68) deletes the *su* in words like *rājā*.

Thus, the *sup* suffixes are an important feature which helps to establish the word boundary of nominals. The feature can be marked on words as shown in the examples below.

1) (राजा(सु[luk,pratyaya])) [...]
2) (विवेक
   ((((: <- र्) उ[lopa,it]) <- स) उ[lopa,it]
)[pratyaya]
) [...]
3) (पर्ण(आत् <- ङसि[pratyaya]))[...]

5 Rules for *ghañ*-ending Words

Words ending in the suffix *ghañ* (*ghañanta*s) are formed by adding the *ghañ* suffix to the verbal roots. The rules A. 3.3.16-55 of the Aṣṭādhyāyī make *ghañanta* words possible. These rules apply the *ghañ* suffix under coocurrence (*upasarga* + *dhātu*) and semantic conditions. For example, the rule *bhāve* (A. 3.3.18) states that the *ghañ* suffix comes after a root to show pure action regardless of doership. The rule *akartari ca kārake saṃjñāyām* (A. 3.3.19) states that it *ghañanta* words indicate appellatives in which an entity is coupled with the action in a non-agentive manner. The rule *parimāṇākhyāyāṃ sarvebhyaḥ* (A. 3.3.20) allows any verbal root to receive the *ghañ* suffix in order to show a measure of quantity.

The following rules control the accent[8] of *ghañanta* words.

1. *ñnityādirnityam* (A. 6.1.197)
   This rule causes words to have acute initial syllables when they are formed using such suffixes that are *ñit* or *nit*. *ghañ* being a *ñit* suffix[9] will make the initial syllable acute unless other rules override the provision of this general rule.

2. *karṣātvato ghaño'nta udāttaḥ* (A. 6.1.159)
   A. 6.1.159 overrides A. 6.1.197 and states that *karṣa* (a *ghañanta* stem derived from the verbal root *√kṛṣã* of the *bhvādi* group, not the *tudādi* one) and other *ghañanta* words that contain *ā* have a acute final syllable. The words formed using the *ghañ* suffix are nominal bases (*prātipadika*s) and get case inflection suffixes.
   Examples:
   a) *karṣá* = *√kṛṣã* (*bhvādi*) + *ghañ*
   b) *pāká* = *√ḍupacaṣ* (*bhvādi*) + *ghañ*

3. *vṛṣādīnāṃ ca* (A. 6.1.203)
   The words that are part of the *vṛṣādigaṇa* have an acute initial syllable. *Vṛṣādigaṇa* contains *ghañanta* words as well as others that are differently derived. The three *ghañanta* words are *kāma*, *yāma* and *pāda*.
   Example:
   pā́da = *√padã* (*divādi*) + *ghañ*, Gaṇa: *vṛṣādi*

4. *tyāgarāgahāsakuhaśvaṭhakrathānām* (A. 6.1.216)
   The words *tyāga*, *rāga*, *hāsa*, *kuha*, *śvaṭha* and *kratha* have an acute initial syllable as an option in addition to their natural accents which happens to be an acute final syllable. Thus, there is an optional accent in the case of *ghañanta*s viz *tyāga*, *rāga*, *hāsa*.
   Example:
   tyāgá, *tyā́ga* = *√tyajã* (*bhvādi*) + *ghañ*.

5. *thāthaghañktājabitrakāṇām* (A. 6.2.144)
   This rule is from the domain of compounds. When a *ghañanta* word appears second in a compound, the rule *gatikārakopapadāt kṛt* (A. 6.2.139) would have retained its default (*prakṛti*) accent. One thing stated in rule A. 6.2.144 is that a *ghañanta* word appearing second in any compound, such as *bhéda* in the tatpuruṣa compound *kāṣṭa-bhedá*, will have an acute final syllable.
   Example:
   kāṣṭhabhedá = *kāṣṭha* + (*√bhidir* + *ghañ*)

6 Outlay of *ghañanta* words

Figure 3 shows a small window in the universe of words. It captures the outlay of all *ghañanta* words according to the rules listed above. The text in italics refers to actual words that the rules treat. Regular text only mentions the categories that can be gathered by a simple analysis of the rule text. The category called *ñidantaḥ* contains such words that end in *ñit* suffixes stated in *ñnityādirnityam* (A.

[8] In showing accent, the grave and acute diacritical marks are used for *svarita* and *udātta* respectively.

[9] The ñ is a marker that gets elided during derivation.

6.1.197). *ghañantaḥ* is a sub-category of *ñidantaḥ* containing such words that end in suffix *ghañ*. The rule *karṣātvato ghaño'nta udāttaḥ* (A. 6.1.159) treats *ghañanta* words having *ā* and the *bhvādi ghañanta karṣa* specially, therefore words from the Having-ā class become another subclass of words within *ghañantaḥ* and *karṣa* is a special *ghañanta* word from the verbal root *√kṛṣã* which appears in the *bhvādi* and *tudādi* compilations of verbal roots (gaṇas). The words *tyāga*, *rāga* and *hāsa* are *ghañanta*s but are specially treated in *tyāgarāgahāsakuhaśvaṭhakrathānām* (A. 6.1.216) so they form a subgroup within the Having-ā class of words in addition to the subgroup of *vṛṣādigaṇa* members *kāma*, *yāma* and *pāda*.

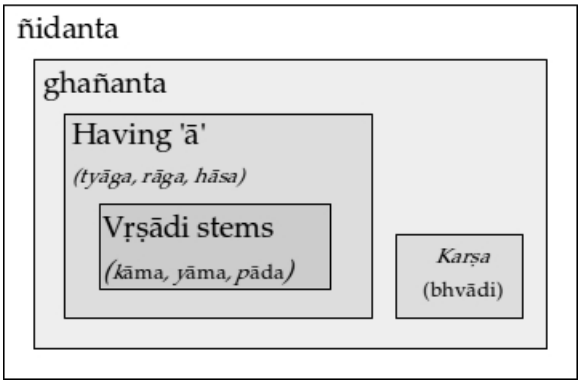

Figure 3: The outlay of *ghañanta* words.

7 Accentuation of *ghañanta*s

Figure 4 shows the accent that each class of *ghañanta*s from figure 3 has. A string of 'F's indicates that words in that class have an acute initial syllable. A string of 'L's indicates that the corresponding words have an acute final syllable. A string of 'O's indicates that the words from that class optionally have an initial or a final acute syllable. A *ñidanta* (ñit-ending word form) has an acute initial syllable due to *ñit* accent of rule *ñnityādirnityam* (A. 6.1.197). A *ghañanta* also has an acute initial syllable because suffix *ghañ* is a *ñit* suffix. The class of *ghañanta*s called 'Having-ā' has an acute final syllable according to *karṣātvato ghaño'nta udāttaḥ* (A. 6.1.159). The specific word *karṣa*, too, is accented by A. 6.1.159 provided that it is not derived from the verbal root *√kṛṣã* belonging to the tudādi group. The members of *vṛṣādi* group of words have an acute initial syllable as stated in *vṛṣādīnāṃ ca* (A. 6.1.203). There are only three *ghañanta* stems viz *kāma*, *yāma* and *pāda* that are members of the *vṛṣādi* group of words. The class called 'Without-ā' contains *ghañanta*s that do not have the syllable *ā*. Finally, the words *tyāga*, *rāga* and *hāsa* are optionally accentuated on the first syllable due to *tyāgarāgahāsakuhaśvaṭhakrathānām* (A. 6.1.216), in addition to being accentuated like the GaYantas in the 'Having-ā' class.

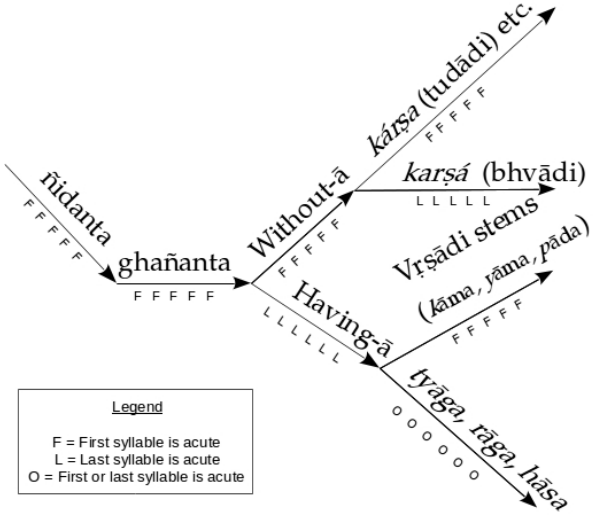

Figure 4: Accents on *ghañanta* words.

8 Features for Accentuation of *ghañanta*s

Section §7 shows how *ghañanta*s are accentuated using Pāṇini's rules. A derivation starts from a set of initial conditions which enjoins the *ghañ* suffix after a verbal root. Over the course of a derivation, some of the rules apply at the right juncture to produce the correct accent. In case of A. 6.1.159, Pāṇini assumes that being *ghañanta* is a feature of a word and that this feature be made known. Similarly, if a *ghañanta* contains '*ā*', that too, is a feature and must be made known so that A. 6.1.159 may apply. Thus, *features* are necessary for proper accentuation. All the features necessary for accentuation of *ghañanta*s are discussed in this section.

In figure 4, each line segment is a feature. Starting from the left hand side, the class *ñidanta* implies that *ñit* is a feature of the word.

The next segment implies that suffix *ghañ* is a feature too. The property of a word to have the vowel *ā* is also a feature. Being a member of the *vṛṣādi* group is also a feature of a word. Since the rules explicitly treat specific words like *karṣa*, *tyāga*, *rāga* and *hāsa* these words also become features by themselves. If all these features are noted on a plain-text word, correct accent can be produced.

Though the set of features discussed above cannot be compromised upon, all the features need not be explicitly noted on the plain text, for some can be deduced. The suffix *ghañ* is a *ñit* suffix so *ñit* feature inheres and need not be explicitly noted. Similarly, we need not explicitly note that the words contain *ā*. Finally, the words *tyāga*, *rāga* and *hāsa* will be dealt with in their respective rules and need not be explicitly mentioned. Since A. 6.1.159 deals with the word *karṣa* which is derived from the verbal root *√kṛṣã* in the bhvādi list, it becomes necessary to include the conjugational class of a verbal root (*dhātugaṅa*) as a feature.

Table 1 summarizes the features that must be noted on *ghañanta*s for accentuation. The contents within braces are meant to suggest an option. The vertical bar separates the alternatives.

SN.	Feature	Rule	Rqd.
1	Suffix *ghañ*	6.1.159	Yes
2	Verb Group	6.1.159	Yes
3	Surface form has ā	6.1.159	No
4	Suffix *ñit*	6.1.197	No
5	*vṛṣādigaṇa* list	6.1.203	Yes
6	{*tyāga* | *rāga* | *hāsa*}	6.1.216	No
7	2nd word in a compound	6.2.144	Yes

Table 1: Features of *ghañanta*s

SN.	Word	Annotation
1	*karṣa*	a) *√kṛṣã* (1) + *ghañ* b) *√kṛṣã* (6) + *ghañ*
2	*pāda*	*√padã* (4) + *ghañ*, *vṛṣādi*
3	*hāsa*	*√hasẽ* (1) + *ghañ*
4	*veda*	*√vidã* (2) + *ghañ*
5	*lekha*	*√likhã* (2) + *ghañ*
6	*kāṣṭhabheda*	*kāṣṭha* + (*√bhidir*(7) + *ghañ*)

Table 2: Examples of annotations on *ghañanta* words

8.1 Represenation of Sense

The rules A. 3.3.18-20 (§5) discussed that *ghañanta*s are formed under three conditions – (i) *bhāve* (pure action), (ii) *saṃjñāyām akartari kārake* (a non-agentive term) and (iii) *parimāṇe* (a measure of quantity). Although, these are semantic conditions of formation and not accentuation of *ghañanta*s, they still need to be represented in the deep structure of *ghañanta*s. The sense that these conditions represent can be referred to by synset IDs from IndoWordNet[10]. The synset ID is a reference number assigned to a set of synonymns. The sense can be annotated as `synset=<SynsetID>` in the deep structure. The senses and synset IDs are discussed below.

bhāve The sense of *bhāva* is represented by the synset ID 12149 which is glossed as *kāraṇasya kārye parivartanasya avasthā* (state in which a cause produces an effect).

saṃjñāyām akartari kārake The word *saṃjñā* can be represented by synset ID 8150 which is glossed as *vyākaraṇaśāstre prayujyamānaḥ saḥ śabdaḥ yaḥ vāstavikaṃ kalpitaṃ vā vastu bodhayati* (that word used in grammar which conveys a real or imaginary thing). For a purpose such as this, the idea of *akartari kārake* can be captured by negating the meaning of the word *kartā* (an agent) represented by the synset ID 11886 and glossed as *vyākaraṇaśāstrānusāreṇa tat kārakaṃ yat dhātvarthasya vyāpārasya āśrayaḥ* (according to grammar, that role, upon which the action of the verb totally depends - an independent agent)

parimāṇe The sense of the word *parimāṇa* is given by synset ID 10743 which is glossed as *bhāraghanaphalādīnāṃ māpanaṃ yena bhavati* (that by which weight, volume etc. are measured)

8.2 Artificial *saṃjñā*s

The `synset` annotation is not a Pāṇinian *saṃjñā*, yet the deep structure can be anno-

[10] http://www.cfilt.iitb.ac.in/indowordnet

tated with it. The term *active*, used to indicate the one active *udātta* among potential ones, is another instance of a non-Pāṇinian *saṃjñā*. It is used in the deep structure of the word *bhaviṣyáti* in Section 3.3. It is not necessary that only formally defined *saṃjñā*s be used in the deep structure. Pāṇini makes use of terms such as *pūrvapadam* and *uttarapadam*, without formally defining them, when discussing *samāsa*s (compounds). A fabricated *saṃjñā* can also be used if it has a use.

8.3 Example of Deep Structure

Table 2 shows a few examples of *ghañanta prātipadika*s (nominal bases). The examples below show deep structure of masculine, nominative singular inflections of some of the words from Table 2.

1. *pādaḥ*
   ((प् (आ[udaatta,active] <- अ) द् अँ[lopa,it]
)[gana=4]

   (घ[lopa,it] अ[udaatta] ञ[lopa,it]
) [pratyaya]

   ((((: <- र्) उ[lopa,it]) <- स्) उ[lopa,it]
)[pratyaya]
) [vrshaadigana, synset=!11886 & 8150]

2. *hāsaḥ*
   Due to optional accent there are two forms – *hā́saḥ* and *hāsáḥ*. The following is the deep structure of *hā́saḥ* produced by the rule *tyāgarāgahāsa...* (A. 6.1.216).

   ((हृ (आ[udaatta,active] <- अ) स् एँ[lopa,it]
)[gana=1]

   (घ[lopa,it] अ[udaatta] ञ[lopa,it]
) [pratyaya]

   ((((: <- र्) उ[lopa,it]) <- स्) उ[lopa,it]
)[pratyaya]
) [synset=12149]

   The deep structure below is for the form *hāsáḥ* produced by the rule *karṣātvato...* (A. 6.1.159).

   ((हृ (आ[udaatta] <- अ) स् एँ[lopa,it]
)[gana=1]

   (घ[lopa,it] अ[udaatta,active] ञ[lopa,it]
) [pratyaya]

   ((((: <- र्) उ[lopa,it]) <- स्) उ[lopa,it]
)[pratyaya]
) [synset=12149]

3. *lekhaḥ*
   ((ल् (ए[udaatta] <- इ) ख् अँ[lopa,it]
)[gana=2]

   (घ[lopa,it] अ[udaatta,active] ञ[lopa,it]
) [pratyaya]

   ((((: <- र्) उ[lopa,it]) <- स्) उ[lopa,it]
)[pratyaya]
) [synset=10743]

9 Deep Morphological Analysis

Sanskrit morphological analyzers are available online on websites of JNU[11] and the Dept. of Sanskrit Studies (Univ. of Hyderabad)[12]. A Sanskrit segmenter is also available at the Sanskrit Heritage Site[13]. Given an unaccented input such as *lābhaḥ*, these morphological analyzers correctly analyse it as a masculine form and produce the stem *lābha* and nominative singular suffix. However, they do not produce a deeper analysis of the stem *lābha* in terms of derivational morphology to show that it is derived from the suffix *ghañ*. Such analysis is required for accentuation of *ghañanta* words. JNU has a separate *kṛdanta* analyzer which gives a deep analysis in terms of the root √*ḍulabhaṣ*, *kṛt* suffix *ghañ* and the nominative singular case.

Some of Pāṇini's rules, such as A. 3.3.18-20 (Section 5), create *kṛdanta*s in specific senses. For semantic processing of *kṛdanta*s which may happen upstream in a stack of NLP applications, it would be good to provide deep analysis of words along with senses that Pāṇini intended. In the case of *ghañanta* words, various features, including the accent, senses and *gaṇa* membership, can be embeded in the deep structure. For example, the word *lābhás* (a

[11] http://sanskrit.jnu.ac.in/morph/analyze.jsp
[12] http://sanskrit.uohyd.ac.in/scl/
[13] http://sanskrit.inria.fr/DICO/reader.fr.html

pre-final form of the *ghañanta* word *lābhaḥ*) is shown below.

```
( ( ङ्[lopa,it]
    उ[lopa,it]
    ल्
    (आ <- अ)[udaatta]
    भ्
    अ[lopa,it]
    ष्[lopa,it]
  ) [dhaatu, group=1]

  ( घ्[lopa,it]  अ[udaatta,active] ञ्[lopa,it]
  ) [pratyaya]

  ( स् उ[lopa,it]
  ) [pratyaya]
 )[synset=!11886 & 8150]
```

The deep structure shown above can be produced by a morphological analyzer created using the Stuttgart Finite State Transducer (SFST[14]). The following snippet[15] shows the FST mapping[16] from the string *lābhas* to the corresponding deep structure. In the mapping various portions of the word, namely *lābh*, *a* and *s*, are mapped to different portions of the deep structure using the language of the SFST (see constructs having pattern x:y). These portions correspond to the basic elements such as the verbal root *√ḍulabhaṣ* from the first (*bhvādi*) group, the suffix *ghañ*, and the nominative singular inflectional suffix *su*.

```
< ( (q[lopa,it]
     u[lopa,it]
     l (A<-a)[udaatta]
     B a[lopa,it]
     z[lopa,it]
    )[dhaatu,group=1]>:{lAB}

 <  (G[lopa,it] a[udaatta,active] Y[lopa,it]
    )[pratyaya]>:a

 <  (s u[lopa,it])[pratyaya]
   )[synset=!11886 & 8150]>:s
```

When a string such as *lābhas* is given to the transducer, the following output is produced which can be parsed as the deep structure of *lābhas*.

```
< ( (q[lopa,it]
     u[lopa,it]
     l (A<-a)[udaatta]
     B a[lopa,it]
     z[lopa,it]
    )[dhaatu,group=1]>

 <  (G[lopa,it] a[udaatta,active] Y[lopa,it]
    )[pratyaya]>

 <  (s u[lopa,it])[pratyaya]
   )[synset=!11886 & 8150]>
```

10 Prioritization of Features

In a traditional derivation, the maxim *paranityāntraghgāpavādānām uttarotaraṃ balīyaḥ* is used to decide which one of the many applicable rules should apply. In the case of accentuation of a ready word, such as *tyāgaḥ*, that opportunity does not arise. Most features in table 1 apply to *tyāgaḥ*. Which rule should be used for accentuation?

The *utsarga-apavāda* nature of rules discussed in section 5 allows a prioritization tree to be created as shown in figure 5. A solid arrow points from a high priority rule to a low priority rule. In the case that a high priority rule applies, it debars all the low priority rules up the chain. For example, in accentuating the compound *kāṣṭhabhedá*, the rule A. 6.2.144 will debar the other rules. A dotted arrow shows an optional rule. The rule A. 6.1.216 is an optional rule. It does not debar A. 6.1.159 but provides an option to it. Accordingly, the words *tyāgaḥ*, *rāga* and *hāsa* will be accentuated in two ways (See example in Section 8.3 item 2).

The priority of the rules from figure 5 allows a prioritization of the features to be created as shown in figure 6. Accordingly, a feature towards the bottom of the figure debars the ones above it when connected using solid arrows. A feature connected using a dotted arrow allow an option. When a high priority feature is found on a word, the corresponding rule noted in the parenthesis should be used for accentuation. In case a high priority feature is not found the next level of features should be checked and so on. If none of the special fea-

[14] http://www.cis.uni-muenchen.de/ schmid/tools /SFST

[15] Snippet is actually a concatenated string, here it is formatted for easy reading.

[16] The mapping uses the Sanskrit Library Phonetic (SLP1) encoding.

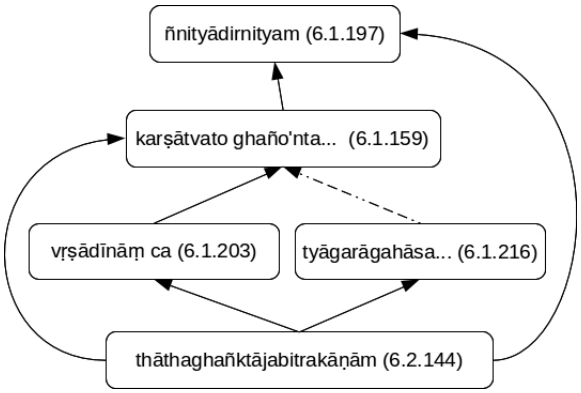

Figure 5: The priority of accent rules for *ghañanta* words

tures exist on a *ghañanta* word, it should be accentuated by the rule *ñnityādirnityam* (A. 6.1.197) due to the feature *ghañ*.

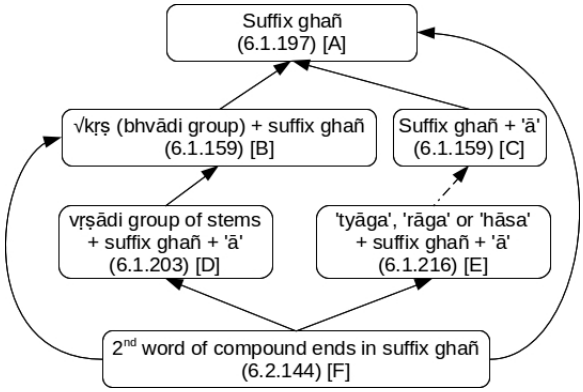

Figure 6: The priority of conditions for accentuation of *ghañanta* words

11 The Feature Arrays

All *ghañanta* words must have one or more primitive features from Table 1. Each of the prioritized conditions in figure 6 refers to a bundle of features. The letter in the square brackets indicates a bundle of features produced on a *ghañanta* word when the corresponding condition is true. A bundle of features can be translated into an array containing serial numbers of the corresponding primitive feature from Table 1. Referring to figure 6, condition [A] can be represented by the array '[1]'. Similarly the condition [E] can be represented by the feature array '[1, 3, 6]'.

Table 3 shows the feature array corresponding to each bundle of features. The ellipsis in the array for cumulative feature [F] is meant to show that other features can coexist but are ignored because only features 1 and 7 are the important in the context of compounds.

The primitive feature numbers may occur only once in the feature array. They need not occur more than once because their repetition will not convey more information than what is conveyed by using them once.

Indicator	Array	Acute Accent On
A	[1]	First syllable
B	[1, 2]	Last syllable
C	[1, 3]	Last syllable
D	[1, 3, 5]	First syllable
E	[1, 3, 6]	First or last syllable
F	[1, 7, ...]	Last syllable

Table 3: Possible feature arrays for *ghañanta* words

12 The Process of Restoration of Accent

A morphological analysis, which containis elements of derivational morphology, is necessary for restoration of accent. A morphological analyzer built on principles discussed in Section 9 will be able to show accent in the deep structure. However, some morphological anaylzers, such as JNU's *kṛdanta* analyzer, produce derivational details but do not provide accent information. Depending upon the availability of accents in morphological analysis, there are two methods to restore accent on plain *ghañanta* words.

In the case that derivational morphology of a word is available without accent information, a computer program can produce a feature array using prioritized conditions shown in Figure 6. Given the feature array, accentuation of the surface form can happen according Table 3. If accent information is available, the accent from the deep structure can be simply copied onto the surface form.

A compound can be split using a Sanskrit compound analyser into *pūrvapada* and *uttarapada* constituents. These words can further be morphologically analysed to check if the second one is a *ghañanta*. For example, *kaṣṭhabheda* can be split into *kaṣṭha* + *bheda*. *bheda* is a *ghañanta*. A program can assigin '[1, 7]' as a feature array to the entire commpound and accentuate it using Table 3.

13 Conclusion and Future Work

The class of *ghañanta prātipadika*s[17] was examined. *utsarga* and *apavāda* rules of accentuation for suffix *ghañ* were studied. At least seven features used to accentuate *ghañanta* words were found. Thus, it is possible to create a set of structural, semantic and accent-related features for a specific morpheme by studying the *utsarga-apavāda* rules related to that morpheme.

Words in Sanskrit can be thought to have a grammatical deep structure, different from the surface form. The deep structure of words contains structural elements like roots and morphological suffixes. It also notes transformations of structural elements leading upto the final form. In addition, it can contain other features of the word such as senses and *gaṇa* membership, if any. Synset IDs from IndoWordNet are a good way to assign a certain sense to words for computational purpose. In this paper, the deep structure for *ghañanta* words was created by looking at some *utsarga-apavāda* rules. A general deep structure can be induced by a thorough study of Pāṇini's rules and how they interact with one another.

The deep structure shows the complete formation of a word, including its accent. Therefore, it can be used to build a morphological analyzer that analyzes the derivational morphology of words in addition to inflectional morphology.

To mark accent on unaccented *ghañanta*s from classical texts, words will have to be morphologically decomposed. Existing morphological analyzers, such as JNU's *kṛdanta* analyzer, can be used or new faster ones can be developed using FST techniques. For restoring accent on compounds having *ghañanta*s for final words, compound analyzers or segmenters like the Sanskrit Heritage Site, will have to be used.

Pāṇinian entities, such as the *vṛśādigaṇa*, can be turned into computational resources for accentution.

References

Kāśīnātha Vāsudevaśāstrī Abhyankara. 2001. *Svaraprakriyā of Rāmacandrapaṇḍita.* Ānandāśrama Publication, Pune, India.

Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Raj Dabre, Archana Amberkar, and Pushpak Bhattacharyya. 2012. Morphological analyzer foraffix stacking languages: A case study of marathi.

G. V. Devasthali. 1967. *Phiṭsūtras of Śāntanava.* Publications of the Centre of Advanced Study in Sanskrit Class C No. 1. University of Poona, Pune, India.

Huet Gérard. 2003. Lexicon-directed segmentation and tagging of sanskrit. In *XIIth World Sanskrit Conference, Helsinki, Finland, Aug*, pages 307–325. Citeseer.

Gérard Huet. 2009. Sanskrit segmentation. *South Asian Languages Analysis Roundtable XXVIII, Denton, Ohio (October 2009)*.

Brahmadatt Jigyasu. 1979. Ashtadhyayi (bhashya) prathamavrtti, three volumes. *Ramlal Kapoor Trust Bahalgadh,(Sonepat, Haryana, India)(In Hindi)*.

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. Introducing sanskrit wordnet. In *Proceedings on the 5th Global Wordnet Conference (GWC 2010), Narosa, Mumbai*, pages 287–294.

Malhar Kulkarni, Samir Sohoni, and Nandini Ghag. 2015. Compsitionality in Pāṇinian grammar. In *Journal Gaurigauravam. Department of Sanskrit, University of Mumbai.*, pages 90–94. Mohiniraj Enterprise.

Peter Scharf, Pawan Goyal, ANUJA AJOTIKAR, and Tanuja Ajotikar. 2015. Voice, preverb, and transitivity restrictions in sanskrit verb use. In *Sanskrit Syntax, Selected papers presented at the seminar on sanskrit syntax and discouse structures*, pages 157–202.

Srisa Chandra Vasu. 1891. *The Ashtadhyayi of Panini. 2 Vols.* Motilal Banarsidass Publishers Private Limited, New Delhi.

[17] *kṛttaddhitasamāsāḥ ca* (A. 1.2.46)