Evaluation of Sankrit Sandhi Tools

Anonymous EACL submission

Abstract

P an . nini s grammar saw the culmination of different thoughts into his monumental work asht adhy ayi . The modern age of information theory has provided a new boost to the studies of asht adhy ay from the perspective of information coding. The importance of asht adhyay is three fold. The first one, as is well known, as an almost exhaustive grammar for any natural language with meticulous details yet small enough to memorize. Though asht adhy ay is written to de- scribe the then prevalent Sanskrit language, it provides a grammatical frame- work which is general enough to analyse other languages as well. This makes the study of asht adhy ay from the point of view of concepts it uses for language analysis important. The third aspect of asht adhy ay is its organization. The set of less than 4000 s utra s is similar to any computer program with one major difference the program being written for a human being and not for a machine thereby allowing some non-formal or semi-formal s utra s which require a hu- man being to interpret and implement them. Nevertheless, we believe that the study of asht adhy ay from programming point of view may lead to a new pro- gramming paradigm because of its rich structure. Possibly these are the reasons, why Gerard Huet feels that Panini should be called as the father of informatics 1 .

The Ashtadhyayi is a list of rules. But these rules, too, are lists: of verbs, of suffixes, and so on. These lists have different headings, and these headings describe the behavior of the items they contain. But the Ashtadhyayi is more complicated than this: ideas in one rule can carry over to the next, or to the next twenty; basic words have specialized meanings; and rules in one chapter may control rules in another. In this way, Panini created a brief and immensely dense work. Thus, we have a large arrangement of different rules that we must try to understand.

The Ashtadhyayi formulated in a morphologically, syntactically, and lexically regimented form of sanskrit. To maximize concision with a minimum of ambiguity, rules are compressed by systematically omitting repeated expressions from them, according to a procedure modeled on natural language syntax (anuvrtti)

It is believed by many scholars that though P an . ini has written a grammar for Sanskrit, the concepts he used are general ones and thus it providesa frame- work to write grammars for other languages.

Sanskrit texts contain numerous words which are formed by the combination of two or more words. This process, known as Sandhi, takes place according to certain rules codified by the grammarian Paanini in his *āstaadhyaayii*. The reverse process of getting back the component words from the Sandhied words is known as Sandhi splitting. This paper attempts to evaluate the performance of the existing Sandhi tools. The evaluation is done both with reference to the rules of Paanini and the words used in actual literature. The performance of the Sandhi tools is found to be terribly bad, and the possible reasons behind this are traced. The word sandhi is an umbrella term that is used to refer to

sound changes that take place when two sounds are close enough. The two sounds may merge to give a single sound, one of the two sounds (the former or the latter) may get changed/reduplicated before combining with the other, or even get elided. A new sound may also come in between.

1 Introduction

āstaadhyaayi (meaning a collection of eight books) by *pāṇini* is the source of Sanskrit grammar, syntax and semantics. The word *sandhi* sandhi "joining", however, does not appear in any of the *sutrā* (rules) of *pāṇini*. There are certain rules that are governed by the condition of *saṃhitā* which as defined in *sutrā* 109 of Chapter 4 of Book 1, means closest proximity of letters. These rules talk about the changes that take place when two letters are in closest proximity. *saṃhitā* is always applicable within a word, between an *upasargā* and a verb root and between words in a compound formation. As regards other cases, it depends on how words are pronounced. If words of a sentence are spoken together without a hiatus, *saṃhitā* applies and sound changes will take place at word boundaries. However, if there is a hiatus between two words, *saṃhitā* does not apply. *sutrā* "formulae/rules" 73-157 of Chapter 1 of Book 6 and all rules of Chapters 3 and 4 of Book 8 describe the changes of sound under the condition of *saṃhitā*. The rules which govern change of sounds when *saṃhitā* is applicable are hereafter referred to as Sandhi (*saṃdhiḥ*) rules.

1.1 Types of Sandhi

- Sandhi can take place either within a word, or between two words. Thus, it is of two types:

1. **Internal Sandhi:** Sanskrit grammar has three kinds of minimal meaningful units (morphemes) prefixes, roots and suffixes and it seems that every word in Sanskrit can be derived from them. When these units combine to form a word, *saṃhitā* applies and certain changes take place. This is known as internal sandhi. For example, *bho* (changed form of verb *bhū* bhuu (to be)) + *anam* (anam, a noun forming suffix) → *bhavanam* (bhavanam, being) is a case of internal sandhi.

2. **External Sandhi:** When sandhi takes place between two words, it is known as external sandhi. For example, *tau* ("both of them") + *ekadā* ("once") → *tāvekadā* ("both of them once") involves external sandhi.

Also, depending on whether the two letters that are combining are vowels, consonants or visarga, sandhis are classified as follows:

1. **Vowel Sandhi:** Both letters are vowels e.g. *hima* ("snow") + *ālayaḥ* ("house") → *himālayaḥ* ("house of snow")

2. **Consonant Sandhi :** At least one of the two letters is a consonant, e.g. *vṛkṣa* ("tree") + *chāyā* ("shade") → *vṛkṣacchāyā* ("shade of tree")

3. **Visarga Sandhi :** A visarga combines with a vowel or a consonant, e.g. *punaḥ* ("again") + *janma* ("birth") → *punarjanma* (punarjanma, rebirth)

1.2 Sandhi Splitting

The process of breaking a sandhied word into the original units it is made of is known as sandhi splitting. A tool which can perform this task is referred to here as a sandhi splitter.

1.3 Sandhi Merging

On the other hand merging of two or more words into a compound word is refereed to Sandhi Merging.

2 Exisiting Sandhi Splitting and Merging tools:

The process of breaking a sandhied word into the original units it is made of is known as sandhi splitting. A tool which can perform this task is referred to here as a sandhi splitter. In the same lines the tools which can merge two words to form the compound word is known as sandhi merging tool. There has been considerable amount of research in the field of sandhi splitting and merging. As of now, three distinct sandhi splitters and merging tools are available:

2.1 Sanskrit Sandhi Analyzer and Splitter

[1] This is a vowel-sandhi splitter. The other two kinds of sandhis (consonant and visarga) are not split by this. This was developed at Jawaharlal Nehru University under the guidance of Prof.

[1] http://sanskrit.jnu.ac.in/rstudents/mphil/sachin.pdf

Girish Nath Jha (Professor, Computational Linguistics, Special Centre for Sanskrit Studies). This is available at http://sanskrit.jnu.ac.in/sandhi/viccheda.jsp.

Architecture of the system is as follows: Pre-processing -¿ Subanta processing -¿ Fixed List Checking -¿ Sandhi Analysis - ¿ Result generator

2.1.1 Pre-processing

Pre-processing includes marking the punctation marks, check the word length to determine whether we can split the words or not, and searching in the corpus if the intput word is already an instance of the examples.

2.1.2 Subanta analysis

This analysis includes splitting the noun phrases into its constituents base (prtipadika) and case terminations (kraka-vacana-vibhakti). Sandhi analyzer will process the base words only.

2.1.3 Fixed List checking

After subanta analysis, the input text has been segmented into base and affixes. Now only base words will be processed for sandhi analysis. But before that these words will be checked into a Dictionary (Monier Williams Sanskrit Digital Dictionary), place name list and noun list. The words found in these resources, will be exempted from sandhi processing. These different files are merged to one text file which is named as lexicon.

2.1.4 Sandhi Analysis

The segmenter class will check the vowel sandhi rule - base in the database to mark the resultant sandhi sounds (marker) for potential splitting and to identify the sandhi patterns for viccheda , corresponding to the marked sound.

2.1.5 Result generator

At each step of marker and pattern identification, the class will check the segmented words in the lexicon to generate the result. For this purpose, it will use Dictionary (Monier Williams Sanskrit Digital Dictionary) and customized Sanskrit corpora as the ling uistic resources. To be a valid segmentation , both the segments must be available in either of the linguistic resources. If the word has more than one sounds marked for sandhi , then only the first word must be present in either of the linguistics resources . The remaining string in this case will continue with the process of rule pattern matching, splitting and search in the linguistic resources.

2.2 Sandhi-Splitter

This was developed at University of Hyderabad under the guidance of Ms. Amba Kulkarni (Associate Professor, Department of Sanskrit Studies). This is available at http://sanskrit.uohyd.ac.in/scl/ . Another Sandhi splitter is available at the TDIL website http://tdil-dc.in/san/sandhi_splitter/index_dit.html but it is the same as the one mentioned above, the only difference being in version. The former is considered here because it is the latest version.

2.2.1 Segmentation Algorithm

The basic outline of the algorithm is(**?**):

1. Recursively break a word at every possible position applying a sandhi rule and generate all possible candidates for the input.
2. Pass the constituents of all the candidates through the morphological analyser.
3. Declare the candidate as a valid candidate, if all its constituents are recognised by the morphological analyser, and all except the last segment are compounding forms.
4. Assign weights to the accepted candidates and sort them based on the weights as defined in the previous subsection.
5. The optimal solution will be the one with the highest weight.

2.3 The Sanskrit Reader Companion

This is a Sasnkrit segmenter and parser, and therefore, also able to split sandhis. This was developed at INRIA, France under the guidance of Professor Gerard Huet, emeritus Professor. This is available at http://sanskrit.inria.fr/DICO/reader.fr.html.

attested preverbs for roots. In a second phase, more stem generation occurs for roots, accounting for the various tenses/moods (lakaras), as well as absolutives and in nitives, but also participles (adjectival kr . dantas) in 10 varieties. Finally, in exional morphology paradigms derive the in ected forms according to the morphological parameters, some of which being read from the lexicon, some of which being de ned in speci c tables. It is not

simple to relate our paradigmatic derivations and the operations of Paninian grammar. Some stem operations relate to morpho-phonetic operations well identi ed in the central rules (stras), such as taking the phonetic grades of gun . a or vr . ddhi . Some are dispersed in various parts of the grammar, such as stem substitutions. Still others, such as retro exion, occur in the nal section (tripadP) of the As . t . adhyayP , which comprises a list of rewrite rules which are applied iteratively at the end of the derivation process, in some kind of phonetic smoothing" phase. Since the forms stored in our data banks are to be matched to the surface realization of the sentence, this phonetic smoothing must be applied at the time of generation of these forms. Some careful analysis must be done in order to understand the mutual interaction of the retro exion rules and of the external sandhi rules (which are used in segmentation in the Heritage reader). This analysis will be given in section 4.4 below. The main discrepancy between the Paninian processes and our reader opera- tions is in the order of rewritings. Often the stras relevant to a given operation are dispersed in di erent sections of the As . t . adhyayP , and thus it is hard to give the trace of the needed stras . However, in many cases, one can recognize the stra operations from the computer program. A complete analysis, in the speci c case of future passive participles

One related diculty concerns the precise de nition of the sandhi relation label- ing implicitly the arcs of the diagram. For most arcs, it is what Western linguists call external sandhi". However, for the arcs issued from the Pv phase, it adds retro exion rules, necessary to explain verbal forms of verbs (or participles) pre- xed by preverb sequences. Furthermore, the forms stored in the various data banks (Noun, etc.) use such retro exion rules in the formation of their stems and of their in exions. The correctness and completeness of our method involves thus a careful as- sessment of the mutual interaction (feeding, bleeding) between the rules of the morpho-phonetic processes of stem formation and in ection, the rules of external sandhi, and the nal tripadP rules.

3 Methodology of Evaluation:

The evaluation was done in two ways: Rule-based and Literature-based.

3.1 Rule-based

The three splitters are evaluated against a corpus which contains at least one example for each of the Paaninian sandhi rules. This brings out how many rules are actually implemented by the splitters. There are 282 examples against 271 rules. This evaluation was done both manually and using a tool developed for this purpose. The corpus is available at ..

For most sandhied words, each of these splitters gives a very large number of possible splits. If any of the splits for a given word matches with the correct split, the splitter is considered to have correctly identified the splits.

While evaluating each of the three splitters for external sandhis, even if the splits are not fully correct and there is some error in the spellings of the words far away from the location where the sandhi takes place, the slightly incorrect split is still considered as correct. An example is *nayanam* (the act of directing) whose correct split is *ne* (changed form of the root nee meaning direct) + *anam* (a noun forming suffix) but *ne* + *anama* is also considered to be correct, even though the last letter does not have a *halanta*. Another example is *prauḍhaḥ*(fully developed, aged, etc) where both *pra* + *ūḍhaḥ* and *pra*+ *ūḍha* are considered correct. However, the automated evaluation rejects the latter result in each of these cases.

This privilege has not been given to internal sandhi cases, which if the splits are only slightly wrong, there are not considered. This is because internal sandhi is between prefixes, roots and suffixes and small mistakes in each of these has the potential to change the meaning. There are some rules which govern combination of letters that may themselves be the results of application of other rules. An example is *vṛkṣas* (vrik.sas, tree) + *śete* ("sleeps") → *vṛkṣaśśete* ("tree sleeps") where the split *vṛkṣaḥ* +*śete* gives the original form. So, even though the corresponding rule has to do with change of *s* to *ś*, the presence of *visarga* is duly considered correct.

Evaluation results are summarized in the following table

There is a significant difference in the results for the UoH and the INRIA splitter in the two modes of evaluation. The results of manual evaluation are detailed below as per the type of the sandhi rules :

Cases not detected by any Splitter- 62 (46.9

Splitting tool	Manual	Automated
JNU	12.4%	11.4%
UoH	26.6%	18.1%
INRIA	19.5%	14.5%

Table 1: Evaluation Results

Splitter	External Sandhi Cases (132)	Internal Sandhi Cases (150)	Overall
JNU	21 (15.9 %)	14 (9.3 %)	12.4 %
UoH	48 (36.4 %)	27 (18 %)	26.6 %
INRIA	49 (37.1 %)	6 (4 %)	19.5 %

Table 2: Evaluation Results

3.1.1 Analysis of Results

- A large number of rules have not been implemented, even if we leave aside those cases where the examples themselves can never be the first two words to start with, i. e. the cases that represent the second or subsequent stages of sandhi between two words, before the final word is obtained.

- The internal sandhi phenomenon seems to have been neglected to a great extent.

Sandhi Merging tool results:

Splitting tool	Manual	Automated
JNU	12.4%	11.4%
UoH	26.6%	18.1%
INRIA	19.5%	14.5%

Table 3: Evaluation Results

3.2 Literature-Based Evaluation:

This takes into account the sandhied words which appear in actual literature. This evaluation is useful because the sandhi splitters are more likely to be used to split such words. If the performance of the splitters in splitting these words is satisfactory, one may consider neglecting the rules which these splitters have not been able to implement, because those rules may not be so frequent in use. However, a bad performance in this context is a cause of actual concern. Five different corpora were used for this purpose:

1. 150-word corpus
2. Manually created Bhagvad Gita corpus containing nine chapters
3. Dictionary-filtered UoH corpora
4. Word-length filtered A.s.taadhyaayii corpus

3.2.1 150-word corpus

This was created from 11 different texts. This has 50 examples from one text, and 10 examples each from the other ten texts. This is smaller in size compared to the other literature corpora, hence the evaluation for first was done both manually and using the automated tool. The other three were evaluated with the help of the tool only because of their much larger size.

Splitter	Manual Evaluation Results	Automated Evaluation Results
JNU	$14/150 * 100 = 9.33\%$	15/150*100=
UoH	$96/150 * 100 = 64\%$	98/150*100=
INRIA	$123/150 * 100 = 82\%$	95/150*100=

Table 4: Evaluation Results

The results of automated evaluation are reported below. The difference of .. between the two sets of results is henceforth considered as the error margin. This is expected to be the boost in performance of the three sandhi splitters if the corpora considered further were to be manually evaluated. Only automated evaluation has been done for them.

3.2.2 Manually created Bhagvad Gita corpus containing nine chapters

The sandhi-split Bhagvad Gita corpus at the UoH website had several limitations which made it unfit to be used for the purpose of automated evaluation. For example, within the first two chapters, out of the total of 431 sandhi cases, there were 41 typos, 92 cases of insufficient splits and 10 cases of even wrong splits.

Thus, a new corpus was manually created. This was done for half of Gita, i.e. for the 9 chapters.

Results of Automated Evaluation:

3.2.3 Dictionary-filtered UoH corpora

The UoH website has 39 sandhi-split corpora but they are not fully correct. There are cases of typos,

	JNU	UoH	INRIA
A_1 **[157]**	2(1.3%)	61(38.8%)	95(60.5%)
A_2 **[270]**	10(3.7%)	115(42.6%)	160 (59.3%)
A_3 **[168]**	11(6.5%)	74(44%)	93(55.4)
A_4 **[164]**	4(2.4%)	78(47.6%)	88(53.7%)
A_5 **[113]**	9(7.9%)	48(42.5%)	60(53.1%)
Total [872]	36 (4.1%)	376(43.1%)	496(56.9%)

Table 5: Evaluation Results

Splitter	No. of Cases Correctly Identified
JNU	3215 (17.5 %)
UoH	11405 (62.2 %)
INRIA	13416 (73.2 %)

Table 6: Evaluation Results

insufficient splits and even wrong splits. Therefore, they were not directly used for the purpose of automated evaluation. The corpora contained thousands of words in total, and a strategy was worked out to create a subset of them that had no errors.

The strategy involved checking whether the splits could be located in some dictionary. Five dictionaries were used for this purpose. We restricted ourselves only to such cases where the splits could be located, even though they may be many cases of correct splits where the splits themselves cannot be located in the dictionary, because of various reasons (dictionaries may not contain all the declensions/ conjugations of a word, nor they are expected to). Further to check that the sandhied words in such cases did not have typos, a sandhi tool was used to do sandhi of the splits identified in the dictionary and check whether the result in each case matched with the sandhied word as given in the corpus. If the two words did not match for a particular case, it was neglected.

Just to illustrate this, we consider the five cases listed below.
tumulo vyanunādayan → tumulaḥ+vi+anunādayan
sarvānbandhūnavasthitān → sarvān + bandhūn + avasthitān
śabda iva → śabdaḥ+ iva
nārhati → na + arhati
astamito bhagavān → astam + itaḥ + bhagavān

The first two cases were not included because at least one word in each of the splits could not be located in any of the dictionaries used for this purpose. The last three cases were considered for the purpose of evaluation.

Results: Total Number of Cases: 18,326

3.2.4 Word-length filtered *āstaadhyaayi* corpus

Many *sutrā* (rules) of *pāṇini* themselves contain many sandhied words. All the sutras with their splits are available at . This was found to be another good source which could be used for the evaluation of the three splitting tools. However, even this source suffered with the limitation of insufficient splits. Moreover, a very significant number of splits could not be expected to be located in any dictionary, because these were the forms of the different sets/ *mahesvara sutrā* that *pāṇini* uses to codify Sanskrit grammar, syntax and semantics. Hence, the strategy used in the previous case to limit to cases of correct splits could not be applied in this case.

Since the problem initially arose because of cases of insufficient splits, another strategy was worked out. The splits which can undergo further splitting themselves are likely to be of greater length than those cases in which further splitting is not possible. The larger the length of the splits, the more likely they are to undergo further splitting. All those cases where the length of the splits was less than a specified word length were analysed together and the results were noted for different values of the word lengths-10,20,30, 40 and 50.

The following five examples are used here for the purpose of illustration. At least one of the splits in each of the first two cases is considerably long, and further splitting is evident. When the word length is reduced, the possibility of further splits is also reduced, though not eliminated. So, in the next two cases, though the word length is reduced, the first split of the third case and the second split of the fourth case can themselves be further split. It is only for the last case that further splitting is not possible.

prathamacaramatayālpārdhakatipayanemāśca → prathamacaramatayālpārdhakatipayanemāḥ+ca
udupadhādbhāvādikarmaṇoranyatarasyām → udupadhāt+bhāvādikarmaṇoḥ+anyatarasyām

taddhitaścāsarvavibhaktiḥ → *taddhitaḥ+ca+asarvavibhaktiḥ*
vṛddhirādaic → *vṛddhiḥ+ādaic*
vija iṭ → *vijaḥ+iṭ*

Results on *āstaadhyaayi* Total Number of Sutras 3,959 Sutras where Sandhi Split Applicable 2,700

No. of letters (≤)	Samples	JNU	UoH
10	93	4(4.3)	21(22.6)
20	571	10(1.75)	100 (17.5)
30	1512	17(1.12)	226(14.9)
40	2045	18 (0.88)	263(12.86)
50	2302	18(0.78)	263(11.42)
All	2700	18 (0.66)	263 (9.74)

Table 7: Evaluation Results

4 Analysis of Errors

Literature-Based Evaluation The literature based evaluation results are also not very impressive. The cases were looked into and the reasons behind the poor performance can be categorised as follows:

4.1 Rules not Implemented

It seems that some of the rules which are even frequent used have not been implemented by one or more of the three splitters. For example, the visarga of *saḥ* and *eṣaḥ* is elided optionally when any letter other than *a* follows it, and there are many such cases of this elision, for example, in Srimad Bhagvad Gita. But none of the three splitters is able to undo this elision to get back the visarga. For example, none of the three splitters is able to do the following split

sa yogī → *saḥ* (that) + *yogī* (who practises yoga)

It will be incorrect to say that rules of elision, in general, are not implemented by any of the three splitters. For example, the split of

bālakā hasanti → *bālakāḥ* (boys) + *hasanti* (laugh) is detected correctly by INRIA.

4.2 Optional Rules

There are some rules which are optional in nature, and less frequently used. For example, when *e* at the end of a pada is followed by a vowel, the *y* of *ay* into which it changes can be optionally elided. The resultant form after elision is less common, but we do have cases in Srimad Bhagvad Gita, for example, of this kind. The following case is an example where none of the three splitters is able to detect the correct split.

vartanta iti → *vartante* (exist) + *iti* (this)

[Had the optional rule not been applied,*vartante +iti* would have led to *vartantayiti*].

4.3 Cascading split effect

There are some rules in which the effect of combination of two words is not restricted to the change in sound at the extreme boundaries of the two words (last sound of first word + first sound of second word). Other letters can also get affected. For example, in

uttara (north) + *ayana* (movement) → *uttarāyaṇa* ("northward movement", refers to movement of Sun towards Tropic of Cancer), the *r* of *uttara* causes the change of *n* of *ayana* into *ṇ*. In absence of *r*, no such change takes place in the case of

dakṣiṇa (south) + *ayana* (movement) → *dakṣiṇāyana* (southward movement, refers to movement of Sun towards Tropic of Capricorn) The three sandhi splitters do not seem to have taken care of such changes. So, while they are able to split *dakṣiṇāyana* correctly, the same is not true for *uttarāyaṇa*.

Another example is the case of change of *s* into *ṣ* when it is immediately preceded by some vowels (for example, *i*, and this *ṣ* changing its subsequent letter because of another sandhi rule, for example *th* into *ṭh*. Thus, we have the examples of:

prati + *sthita* → *pratiṣṭhita* ("well-established")

and *yudhi* + *sthiraḥ* → *yudhiṣṭhiraḥ* (one who is stable in war) where the sandhied forms are **not split by any of the three splitters**.

4.4 Multiple splits

The process of sandhi splitting involves splitting the sandhied word at different potential locations, and validating the splits to check which one of them is correct. If the set used for validation is not complete, even correct splits may sometimes not be validated. For example, in

a (not) + *chedyaḥ* ("solvable", "penetrable") → *acchedyaḥ* ("not solvable/penetrable")

the fact that none of the three splitters has been able to split the sandhied word may have to do with the possiblity that may *a* not have been validated as a proper split.

4.5 Compounding effect

The process of compounding, due to which words come together without necessarily their being a change when they merge, also creates problems. While the UoH and the INRIA tools do have the provision of decompounding along with sandhi splitting, the JNU splitter does not have a way to do both together. For example,

lakṣyasyārthatvavyavahārānurodhena → *lakṣyasya + arthatvavyavahāra + anurodhena*

The second split is not validated without decompounding, and thus even though, only vowel sandhis are involved, the JNU splitter is not able to correctly split the word. Even the INRIA and UoH splitters are not always able to get around this problem. For example, none of the three splitters is able to detect this:

prapañce'vāntaravibhāgapravibhāgabhinnānantapadārthasaṅkule'pi → *prapañce + ava + antara-vibhāga-pravibhāga-bhinna + ananta-pada + artha-saṅkule + api*