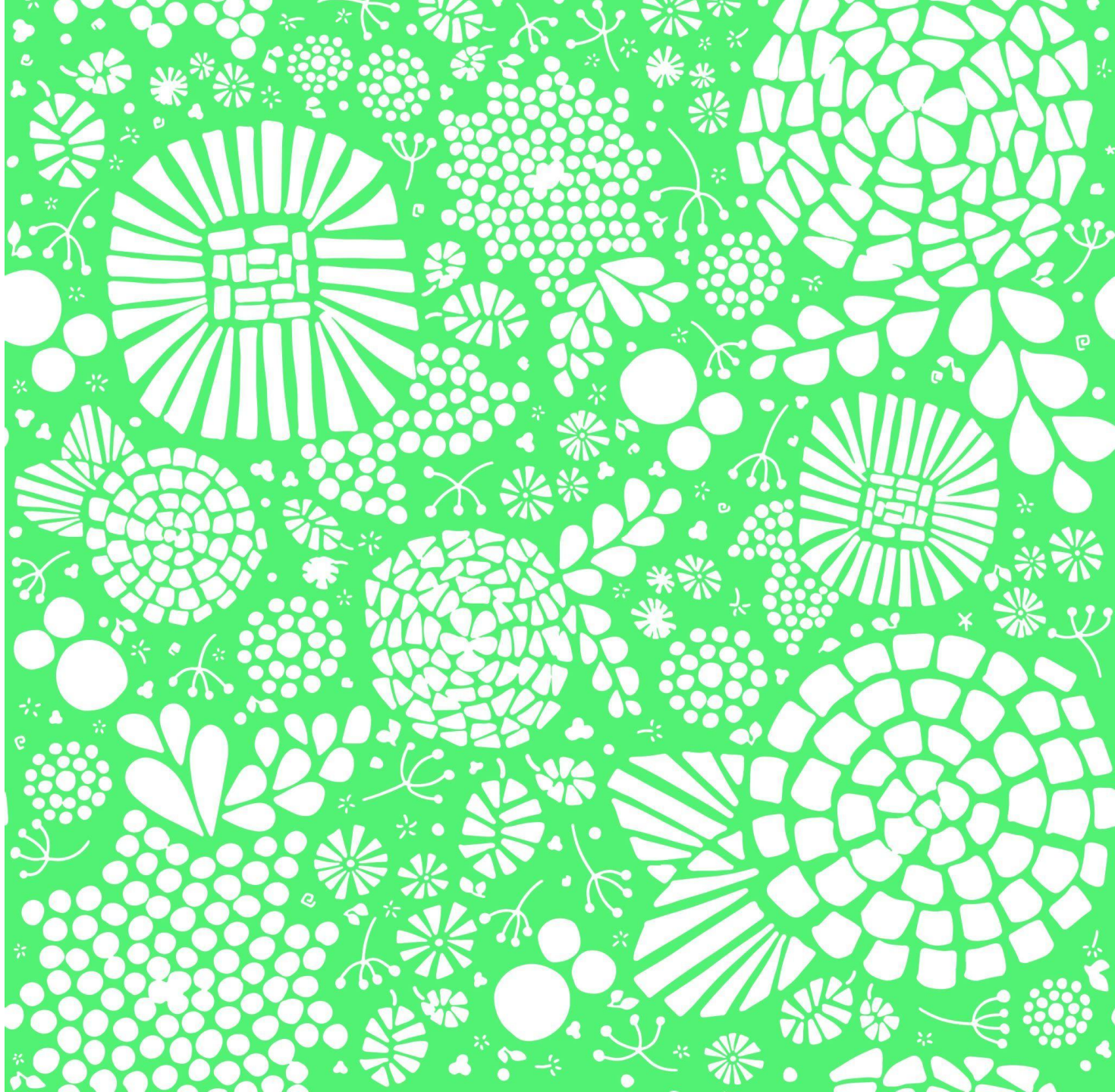
BharatGPT

MULTI-LINGUAL TEAM

Badri Vishal Kasub a

Weekly Updates



[18th Feb – 25th Feb] Task updates

- **Assigned with task of Pre-tokenization : SANDHI SPLITTING**

- Learnt the pipeline of tokenization phase - flowchart
- Requirement of efficient Sandhi-splitter model to pre-tokenize words
- Implemented 3 opensource repositories for Sandhi splitting
- Ideas for efficient Sandhi Splitting have been proposed
- Discussed with the team to address some existing sandhi-split challenges

Sandhi Splitter Models

- Malayalam only Sandhi splitter - link
  - Trained on nearly 4k sandhis, model training code is available
  - They inform that it can be enhanced to various languages
  - Implemented the splitter and works decent for malayalam and not for other languages
  - Work was done in 2016, very old

- Sandhi Prakarna [for Devanagari] repo - link
  - A seq-to-seq training model for Sandhi vicched is available
  - Have trained the model for 117k sandhis and test on 4k sandhi words, works decently
  - The work could be utilized for training sandhi splitter

- Kannada Sandhi spliiter - link
  - Work was done on 2017, and could be used for kannada sandhi splitting

Samsadhini – Sanskrit Heritage Reader(SHR)

- Assigned with task of Implementing Sandhi Splitter of **Samsadhini**
- Initially, got guided to install entire SHR application from here, then later got into sandhi-splitter repository of SHR : GitHub repo
- Plan is to install sandhi splitter **locally,** so to avoid requirement **to call api** for large data of words that we have for BharatGPT

Samsadhini – Sanskrit Heritage Reader(SHR)

- **Installed and checked working model of Samsadhini Sandhi – splitter locally**

- We get list of possible segments, morphology of each segment that is predicted.

- It has support for sanskrit sandhi split for devanagari, english scripts


```
"morph": [
    {
        "word": "विद्या-",
        "stem": "विद्या",
        "root": "",
        "derivational_morph": "",
        "inflectional_morphs": [
            "iic."
        ]
    },
    {
        "word": "वित्-",
        "stem": "विद्#3",
        "root": "",
        "derivational_morph": "",
        "inflectional_morphs": [
            "iic."
        ]
    },
```



```
"input": "विद्यालय",
"status": "success",
"segmentation": [
    "विद्या-लय",
    "विद्या-आलय",
    "विद्या लय",
    "विद्या आलय",
    "विदी आलय",
    "विदि आलय",
    "वित्-य-आलय",
    "विद्या-आल-य",
    "विद्या-अल-य",
    "विद्या आल-य"
],
```

Some points

- Existing Sandhi Splitter has issues , have to address that like Vidhyalay
- Though the Sanskrit Sandhi Splitter of Samsadhini is good and there can be rule based post-processing to check for correct sandhi split validation
- There are defined rules for sandhi merge, so dataset could be created through sandhi merge rules, and then a model could be trained for sandhi split (applicable for sanskrit, devanagari)
- For other Indic languages, Sandhi splitter availability is few

Some Questions

- When we have large dataset, **does splitting Sandhi matters?**

- Since BPE tokenization would be used, we anyhow split char level
- ["hug", "pug", "pun", "bun", "hugs"]
- ['h', 'u', 'g', 'p', 'u', 'n', 'b', 'u', 'n', 'h', 'u', 'g', 's']

- ["रामः गच्छति", "विद्यालय", "विद्यालयम्", "विद्यालयात्", पाठशाला]
- ['रा', 'मः', ' ', 'ग', 'च', 'छ', 'ति', ' ', 'वि', 'द', 'या', 'ल', 'य', ' ', 'वि', 'द', 'या', 'ल', 'य', 'म्', ' ', 'वि', 'द', 'या', 'ल', 'य', 'ा', 'त्', ' ', 'पा', 'ठ', 'श', 'ा', 'ल', 'ा']

- When data is huge, does sandhi splitting matters cause BPE works better when data is huge

References

- https://arxiv.org/pdf/2010.12940v1.pdf
- https://github.com/SushantDave/Sandhi_Prakarana
- https://github.com/libindic/sandhi-splitter
- https://towardsdatascience.com/byte-pair-encoding-subword-based-tokenization-algorithm-77828a70bee0
- https://github.com/SriramKrishnan8/sandhi_vicchedika

Running Notes

- Bigger ngram and largest frequency, completely statistical, splits very minutely
- YOU Have to work on SANDHI Spliiter [TASK_TO_DO]
- Existing one is we have efficient merging sandhi, but for splitting multiple possible outcomes, that is the problem [TASK TO DO]
- A tokeniser is better if no. of splits are less [LOGIC why]
- Training a tokeniser to get a sandhi splitter [END GOAL]
- Fertility scores [READ]
- How good is a tokeniser [READ]
- Statistical splitter from IndicNLP [Viterbi also + READ]
- Keep stem of the word if it is statistical. For Vidyalaya case, probability of ALAYA is more than LAYA [IDEA to LOOK INTO]
- IS BPE not enough?? [COMPARE WITH FERETILITY SCORE so we know about it]

Sandhi Examples

రాముడు + అతడు = రాముడతడు

Rama Him He is Ram

राम वह राम + वह

रमुद्दु अथद्दु रमुदथद्दु

TELUGU SAMPLE

నాడు + నాడు = నానాడు

నాడు నాడు

Transliteration might not be helpful for all cases for splitting in various languages like Telugu

Majority of sandhi words are in sanskrit, but not all of them

[26th Feb – 03rd Mar] Task updates

- Midsems week, not much work done
- Exams on 29th Feb and 02nd March

[04th Mar – 10th Mar] Task updates

- Processing and providing results of Sandhi vicced from Samanthar dataset
- Added documented Code into GitHub - link
- Comparision of BPE tokenizer and Sandhi splitting + BPE

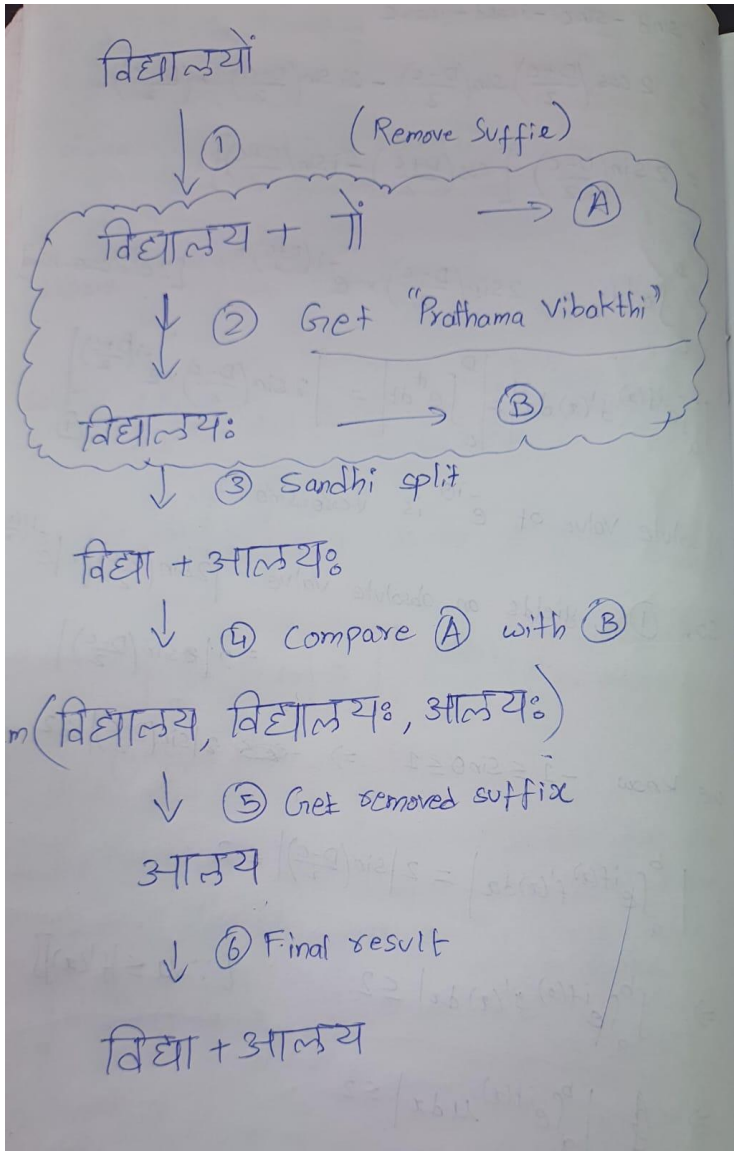
Adding Prathama Vibhakthi

- **Vibhakthis are not present for all the words**
- **API Call through SHR local installation to get the vibhaktis**

http://10.198.63.39/cgi-bin/SKT/sktdeclin?lex=SH&q=vidyalaya&t=VH&g=Mas&font=deva

Word	Vibhakti
अस्ति	अस्तिः
विभाज्	विभाक्
तुर्	तूः
कथ्	कथ्
वर्ण्	वर्ण्
प्रशास्	
उन्मूल्	उन्मूल्
परिवह्	परिवाट्

Missing cases for Masculine and Feminine based words for Vibhaktis



Adding Prathama Vibhakthi

aa
ii
uu
.rr
e
ai
o
au

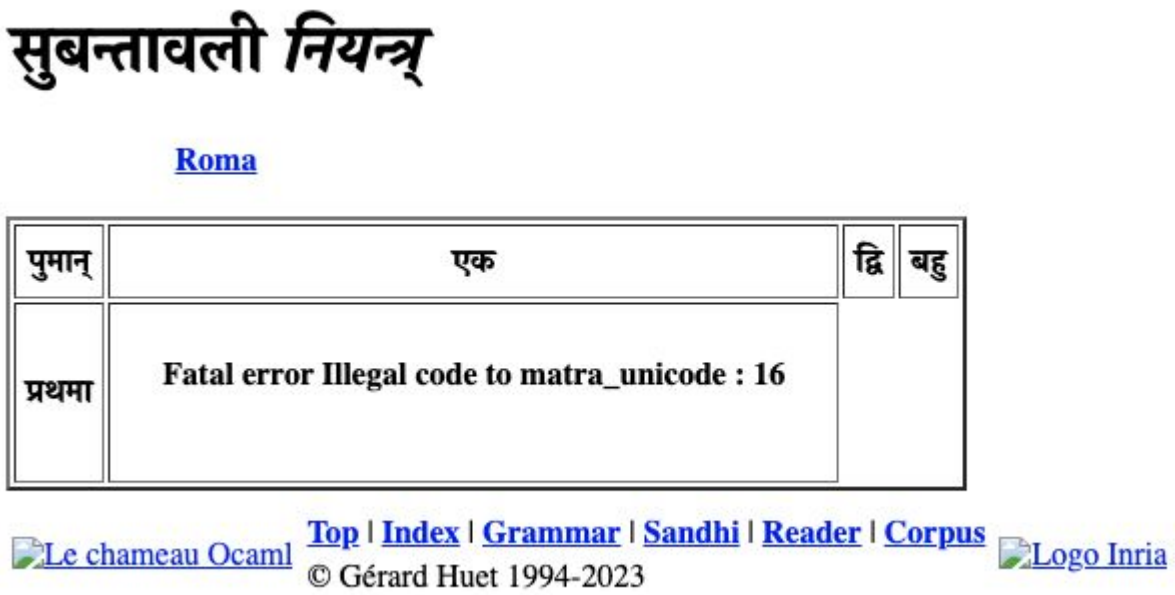
Feminine Type ending suffixes

1. Check for Masculine or Feminine word type
2. Run API call to get vibhaktis based on word type
3. Take the prathama vibhakti from the vibhakti table
4. If vibhakti is not found, check for another word type
5. Else the same input word is passed at the end

It is taking 2hrs to fetch vibhaktis for 1.8 lakh words since it is API call
it is taking 43hrs to get the sandhi vicched output for 1.8 lakh words

For stem words out of 1667, 320 errors were encountered. 2 types of errors are there

Issues Faced



सुबन्तावली *नियन्त्र्*

Roma

पुमान्	एक	द्वि	बहु
प्रथमा	Fatal error Illegal code to matra_unicode : 16		

Le chameau Ocaml Top | Index | Grammar | Sandhi | Reader | Corpus Logo Inria
© Gérard Huet 1994-2023

1. No Such matra word exists

2. Words where No Vibhakthi exists for both Masculine and Feminine types

Adding Prathama Vibhakthi

Original	Sandhi split
अस्ति	अस्ति
विभाज्	#विभाज्
तुर्	#तुर्
कथ्	#कथ्
वर्ण्s	#वर्ण्स्
विद्याल	विद्या-
य	लय

Original	Vibhakti	Sandhi split
अस्ति	अस्तिः	#अस्तिः
विभाज्	विभाक्	#विभाक्
तुर्	तूः	#तूः
कथ्	कथ्	#कथ्
वर्ण्s	No response	No response
विद्याल	विद्यालयः	विद्या-आलयः
य		

अस्ति	अस्तिः	#अस्तिः
विभाज्	विभाक्	#विभाक्
तुर्	तूः	#तूः
कथ्	कथ्	#कथ्
वर्ण्s	वर्ण्s	#वर्ण्स्
विद्याल	विद्यालयः	विद्या-आलयः
य		

Processing Step	Words
Original File	180431
Unique words	24827
Words having more than 2 vowels	18020

Evaluation Metrics

1) Fertility : Avg Number of splits per word

2) Continuation Rate : Avg number of words getting split

Variation 1:
Using Pratyay split with PyCDSL dictionaries,
Prathama Vibhakthi of stem words before Sandhi Splitting

Processing Step	Words
Words that were split	5240
Unique words	19757
Continuation Rate	0.265
Total number of splits	23295
Fertility	1.179

Meeting Notes

Comparision should be like this

1. Ideal: Without any pretokenization -> BPE -> continuation and fertility computation, vocabulary size
2. Compare : With Sandhi, - > BPE -> continuation and fertility compuation, vocabulary size

3. #of words, should be after sandhi - >
4. Only unique words or all the words for comparision

BPE vs Sandhi splitting on Samantar data

IDEAL [BPE Tokenization]

- Vocab size : 50k
- Fertility score: 1.4

Pratyaya + Sandhi Splitting + BPE Tokenization

Processing Step	Words
Words that were split	132335
Total words	182097
Continuation Rate	0.265
Vocab size	21649
Fertility score	1.09576

Variation 1: Using Pratyay split with PyCDSL dictionaries, Prathama Vibhakthi of stem words before Sandhi Splitting

ALL WORDS

Processing Step	Words
Words that were split	132335
Total words	182097
Continuation Rate	0.265
Total number of splits	21649
Fertility	1.09576

[11th Mar – 17th Mar] Task updates

- Sandhi words pipeline complete setup
- Providing sandhi splits in csv format
- Handling of all possible values of word in well documented json file with storing all word information
- Optimizing the issues faced in sandhi splitting and pipeline
- Vibhakti to word conversion
- Fetching the top-k splits of Vibhaktis

- Attended talk with Professors on discussion of Sandhi splits

Process of Sandhi Splitting

1. Remove all punctuations from txt file
2. Count the frequency and find unique words
3. Collect stem and suffix words if pratyaya exists
4. Prathama Vibhakti Identification
5. Sandhi Split on Word and Pratama Vibhakti of the word
6. Storing the results in Json file

Postprocessing of Results

- Removal of Vibhakti sandhi-split suffix to original word
- Converting the json results into required CSV files
- Code to have the stem, suffix, top-k sandhi-splits into the csv file

व्यवहार	व्यवहार	
सन्दर्भ	सन्दर्भ	
पहल	पहल	
महासचिव	महा	सचिव
पडे	पडे	
सहमति	सह	मति
नये	नये	
विकल्प	विकल्प	
करीना	करीना	
चण्डीगढ	चण्डीगढ	
आठ	आठ	
पारित	पारित	
की,	की,	
कार्यों	कार्यों	
हिरासत	हिरासत	
देर	देर	
पालन	पाल्	अन
पाक	पाक	
बढोतरी	बढोतरी	
निपटने	निपटने	

[18th Mar – 24th Mar] Task updates

- **Sandhi splits on IndicCorp data for Top 10k words for Hindi data are provided**
- **For Multi-lingual data, Transliteration of words and Hindi sandhi-splitting is not ideal approach**

రాముడతడు = రాముడు + అతడు

रमुदथदु = [#रमुदथदु]

- **So 1 sandhi splitter model is not good enough for multi-lingual sandhi splitting**
- **Sandhi splitters are not openly available for all Indic languages like Tamil or Telugu, have to create one.**
- Currently provided for 10k sandhi splits, without upasarga removal

Telugu Sandhi Formation Rules are well defined

Common Sandhi Split Rules same as Sanskrit Language

S.No	Telugu Word	English
1	సవర్ణదీర్ఘ సంధి	Suvarna Dheerga Sandhi
2	గుణ సంధి	Guna Sandhi
3	వృద్ధి సంధి	Vrudhi Sandhi
4	యణాదేశ సంధి	Yanadesha Sandhi
5	జశ్త్వ సంధి	Jashtwa Sandhi
6	శ్చుత్వ సంధి	Schutva Sandhi
7	అనునాసిక సంధి	Anunasika Sandhi
8	విసర్గ సంధి	Visarga Sandhi
9	పరసవర్ణ సంధి	Parsavarna Sandhi
10	పరరూప సంధి	Pararoopa Sandhi

S.No	Telugu Sandhi Rules	English Translated
1-3	అత్వ, ఇత్వ, ఉత్వ సంధి	Athva, Ethva, Uthva Sandhi
4	యడాగమ సంధి	Yadagama Sandhi
5	టుగాగమ సంధి	Tugagama Sandhi
6	రుగాగమ సంధి	Rugagama Sandhi
7	దుగాగమ సంధి	Dugagama Sandhi
8	నుగాగమ సంధి	Nugagama Sandhi
9	ద్విరుక్తటకార సంధి	Dvirukta takara Sandhi
10	సరళాదేశ సంధి	Saraladesha Sandhi
11	గ, స, డ, ద, వా దేశ సంధి	Ga, Sa, Da, Dha, Va, adesha Sandhi
12	ఆమ్రేడిత సంధి	Aamreditha Sandhi
13	పుంప్వాదేశ సంధి	Pumpvadesha Sandhi
14	త్రిక సంధి	Thrika Sandhi
15	పడ్వాది సంధి	Padvadhi Sandhi
16	ప్రాతాది సంధి	Prathadi Sandhi
17	లు, ల, నల సంధి	Lu, La, Na Sandhi

IndicCorp Sandhi split results

- Total unique words in IndicCorp : 20,757,052 (20 million words)
- Cleaned unique words list of IndicCorp : 10,264,219 (10 million words)
- Provided cleaned words with the frequency of their occurrence
- Considering words with more than 2 vowels only to Sandhi Split
- Code to generate results end to end of Sandhi splits with upsarga and pratyaya removal for 99,998 words

Why are we doing Sandhi splitting?

- What , Why and How
- What - > IT is answered that, we want to split complex words into simple words
- Why -> Why are we doing Sandhi Splitting
- we want to train for setting up better tokenization pipeline
- How -> How is this going to be useful

[25th Mar – 31st Mar] Task updates

- Generation of Sandhi Splits for **2 lakh processed words** from IndicCorp
- Storing the sandhi splits into json file


```
"word": "प्रबन्धन",
"frequency": 792357,
"transliterated": "prabandhana",
"gender": "Fem",
"prefix_exists": true,
"suffix_exists": true,
"prefix": "प्र",
"stem": "बन्ध्",
"suffix": "अन",
"vibhakti": null,
"I_velthuis": "bandh",
"sandhi_split_happens": false,
"sandhi_split": null
```


Added all possible parameters for the sandhi split words as required

1. Prefix Split
2. Suffix Split
3. Vibhakti fo the Stem
4. Sandhi Split of the Stem
5. Sandhi Split of the Vibhakti
6. Processing Vibhakti Sandhi Split

Fetching stem by removing Suffixes

- Suffix list: ['ते', 'ई', 'एगी', 'ना', 'यें', 'ता', 'ती', 'ऊंगी', 'जिये', 'एं', 'येंगी', 'येंगे', 'ऊंगा', 'आ', 'कर', 'येगी', 'ऊं', 'ओगे', 'एंगे', 'एंगी', 'ओ', 'ए', 'ये', 'यी', 'ने', 'एगा', 'ओगी', 'इये', 'येगा', 'या'] fetched from link
- Based on repeatance of suffixes from stem words, split the suffixes with the stem words
- Example:-
- फँसा,15, ['ए', 'ओ', 'ये', 'एं', 'यें', 'ने', 'एगी', 'कर', 'एगा', 'या', 'ते', 'ई', 'ता', 'ना', 'ती']
- Stem: फँसा

[01st Apr – 07th Apr] Task updates

1. Updated the manual code of fetching stems from suffix removal, **along with frequencies**
2. 373 words from 2lakh processed words had stems, whose suffix>3
3. 18,800 stem words from 10lakh cleaned hindi words of IndicCorp

- Updated the sandhi split code with the better suffix split code

- Extracted valid sandhi splits from the 1.7lakh words : 28,310 splits

Tokenization with Sandhi Splits

- Collected Samanatar dataset of 10k sentences: 182201 words
- Out of 28,310 valid sandhi splits, Samanatar dataset had 2491 words where sandhi split happened
- The sandhi split resulted in increase of 2582 words. (avg split:2.03 per word)

Datafile name	Words
Samanantar	182201
Processed Samanantar	184783

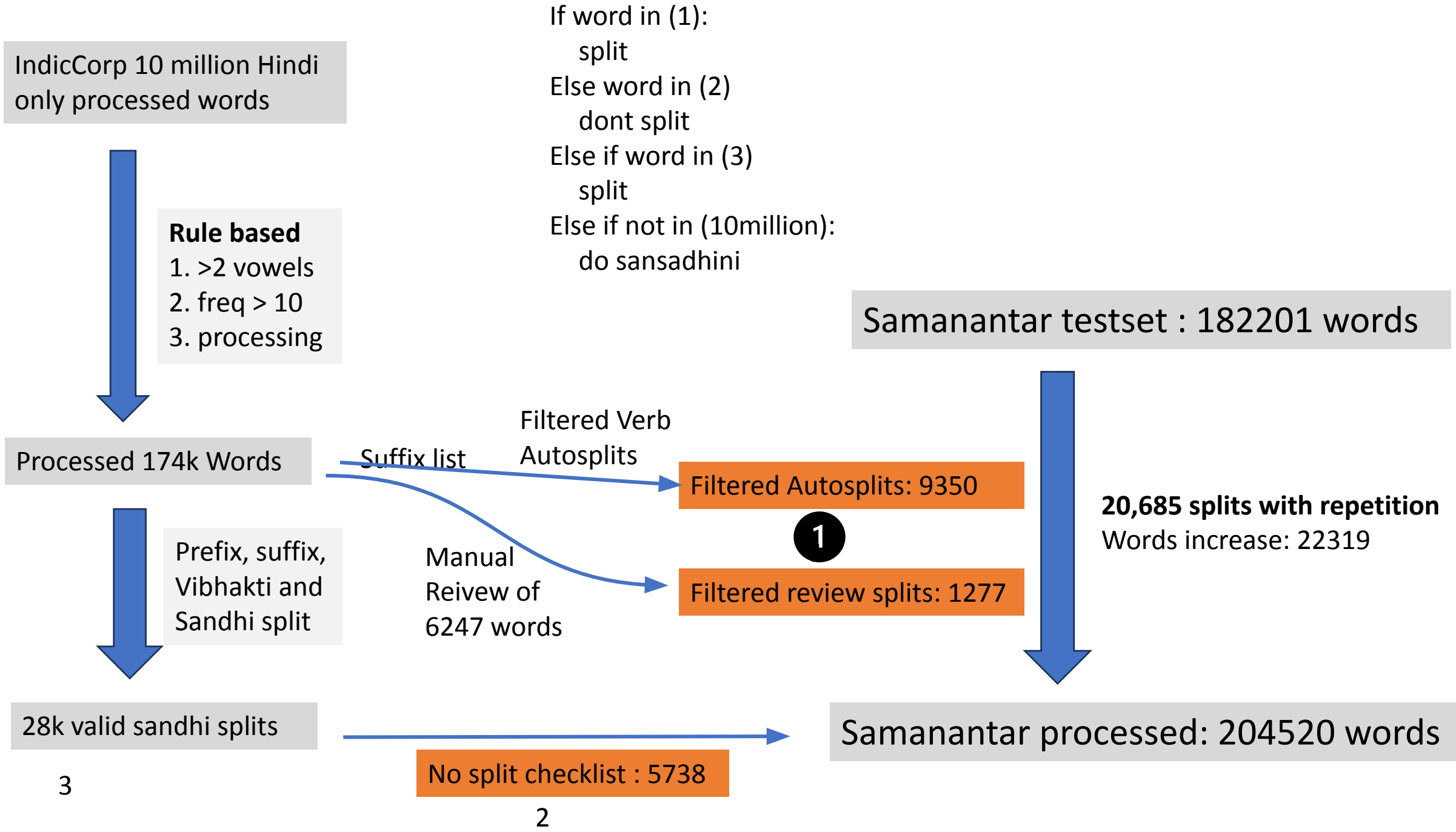
Tokenization with Sandhi Splits

- Collected Samanatar dataset of 10k sentences: 182201 words
- Out of 28,310 valid sandhi splits, Samanatar dataset had 2491 words where sandhi split happened
- The sandhi split resulted in increase of 2582 words. (avg split:2.03 per word)

Datafile name	Words
Samanantar	182201
Processed Samanantar	184783

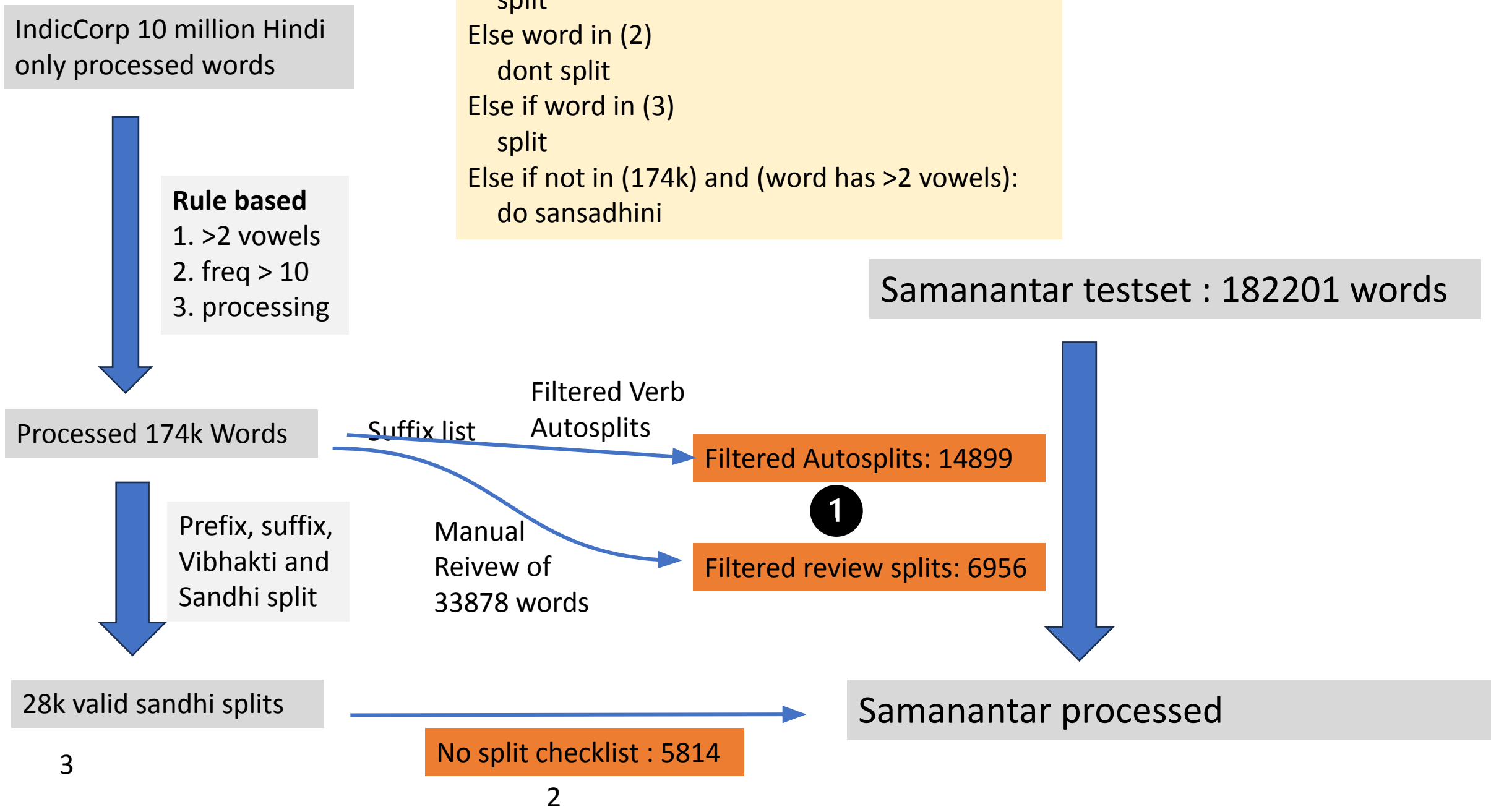
[08th Apr – 14th Apr] Task updates

1. **Pipeline setup for Training BPE Tokenization to use Sandhi-splits**


If word in (1):
split
Else word in (2)
dont split
Else if word in (3)
split
Else if not in (10million):
do sansadhini
IndicCorp 10 million Hindi only processed words
Rule based
1. >2 vowels
2. freq > 10
3. processing
Processed 174k Words
Suffix list
Filtered Verb Autosplits
Filtered Autosplits: 9350
1
Manual Reivew of 6247 words
Filtered review splits: 1277
Samanantar testset : 182201 words
20,685 splits with repetition
Words increase: 22319
Prefix, suffix, Vibhakti and Sandhi split
28k valid sandhi splits
3
No split checklist : 5738
2
Samanantar processed: 204520 words

[15th Apr – 21st Apr] Task updates

1. **Pipeline setup for Training BPE Tokenization to use Sandhi-splits**


If word in (1):
split
Else word in (2)
dont split
Else if word in (3)
split
Else if not in (174k) and (word has >2 vowels):
do sansadhini
IndicCorp 10 million Hindi only processed words
Rule based
1. >2 vowels
2. freq > 10
3. processing
Processed 174k Words
Suffix list
Filtered Verb Autosplits
Filtered Autosplits: 14899
1
Manual Reivew of 33878 words
Filtered review splits: 6956
Samanantar testset : 182201 words
Prefix, suffix, Vibhakti and Sandhi split
28k valid sandhi splits
3
No split checklist : 5814
2
Samanantar processed

Telugu Sandhi Formation Rules are well defined

Sanskrit Language inspired Sandhi Splits

S.No	Telugu Word	English
1	సవర్ణదీర్ఘ సంధి	Suvarna Dheerga Sandhi
2	గుణ సంధి	Guna Sandhi
3	వృద్ధి సంధి	Vrudhi Sandhi
4	యణాదేశ సంధి	Yanadesha Sandhi
5	జశ్త్వ సంధి	Jashtwa Sandhi
6	శ్చుత్వ సంధి	Schutva Sandhi
7	అనునాసిక సంధి	Anunasika Sandhi
8	విసర్గ సంధి	Visarga Sandhi
9	పరసవర్ణ సంధి	Parsavarna Sandhi
10	పరరూప సంధి	Pararoopa Sandhi

S.No	Telugu Word	English
1-3	అత్వ, ఇత్వ, ఉత్వ సంధి	Athva, Ethva, Uthva Sandhi
4	యడాగమ సంధి	Yadagama Sandhi
5	టుగాగమ సంధి	Tugagama Sandhi
6	రుగాగమ సంధి	Rugagama Sandhi
7	దుగాగమ సంధి	Dugagama Sandhi
8	నుగాగమ సంధి	Nugagama Sandhi
9	ద్విరుక్తటకార సంధి	Dvirukta takara Sandhi
10	సరళాదేశ సంధి	Saraladesha Sandhi
11	గ, స, డ, ద, వా దేశ సంధి	Ga, Sa, Da, Dha, Va, adesha Sandhi
12	ఆమ్రేడిత సంధి	Aamreditha Sandhi
13	పుంప్వాదేశ సంధి	Pumpvadesha Sandhi
14	త్రిక సంధి	Thrika Sandhi
15	పడ్వాది సంధి	Padvadhi Sandhi
16	ప్రాతాది సంధి	Prathadi Sandhi
17	లు, ల, నల సంధి	Lu, La, Na Sandhi

Telugu Efficient Sandhi Splitting

- Indic Corp -> Translate -> Sandhi-split Sansadhini -> Splits in Sanskrit -> convert to Telugu