Morphological Analysis and Generation for Pali

David Alfter

Jürgen Knauth

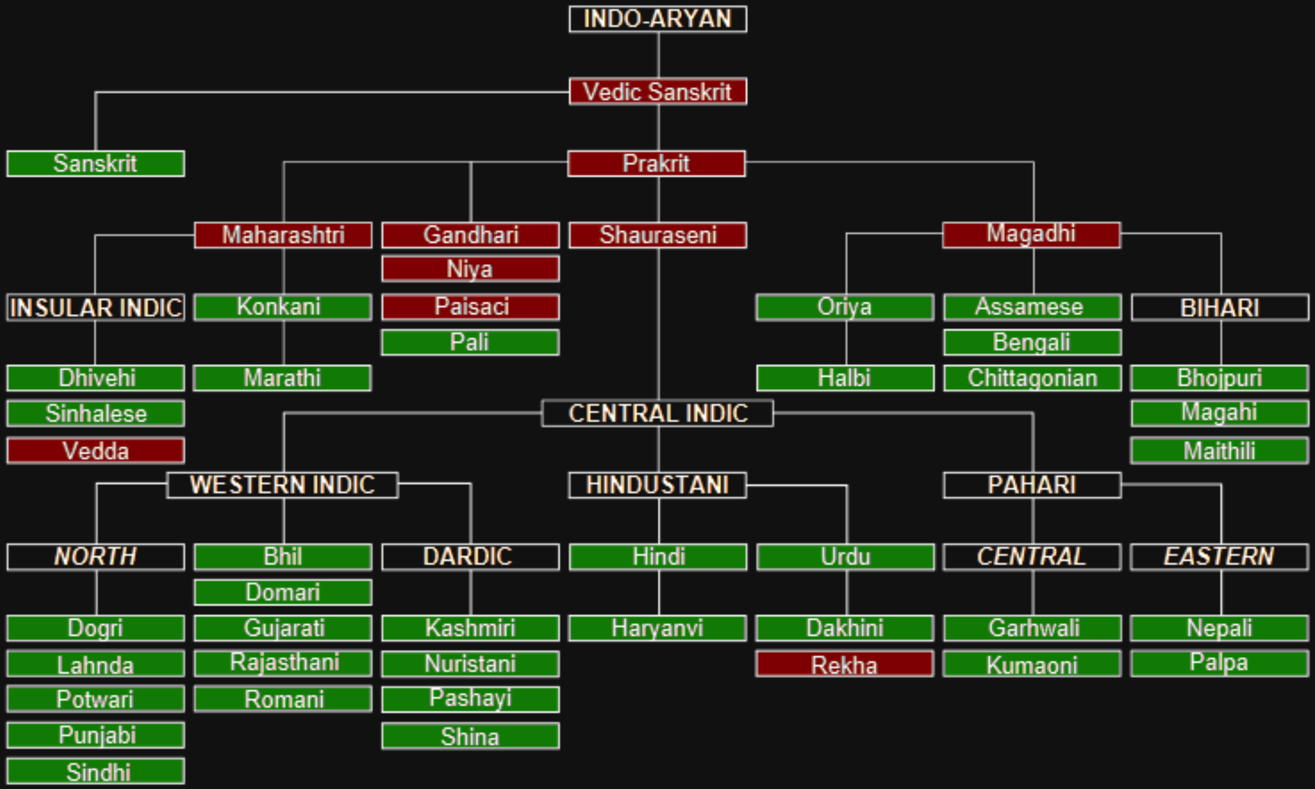
18 September 2015

@daalft

Pali

Pali

- (Dead) Indo-aryan language
- Fusional language
- Rich morphology
- Sandhi



Source:
https://commons.wikimedia.org/wiki/File:BoreanLanguageTree.png

Fusional language

Morphological information added by affigation

No 1:1 correspondence

DEVO

- Base: DEV-
  - god / deity
- Ending: -O
  - noun
  - singular
  - masculine
  - nominative

Compounding

naccagītavāditavisūkadassanamālāgandhavilepanadhāraṇamaṇḍanavibhūsanaṭṭhānā

Compounding

naccagītavāditavisūkadassanamālāgandhavilepanadhāraṇamaṇḍanavibhūsanaṭṭhānā

dancing singing music show-watching garland perfume cosmetics wearing decoration decoration

Compounding

naccagītavāditavisūkadassanamālāgandhavilepanadhāraṇamaṇḍanavibhūsanaṭṭhānā

dancing, singing, music, going to see entertainments, wearing garlands, using perfumes, and beautifying the body with cosmetics

7th precept

naccagītavāditavisūkadassanamālāgandhavilepanadhāraṇamaṇḍana
vibhūsanaṭṭhānā veramaṇi sikkhāpadaṃ samādiyāmi

I adopt the precept of refraining from ...

Sandhi

External sandhi

evaṃ ca (and thus) → evañca

Internal sandhi

paca + ti → pacati (he cooks)

paca + mi → pacāmi (I cook)

canda (moon) + udayo (rising) → candodayo (rising of the moon)

Internal sandhi

paca + ti → pacati (he cooks)

paca + mi → pacāmi (I cook)

canda (moon) + udayo (rising) → candodayo (rising of the moon)

The Problem

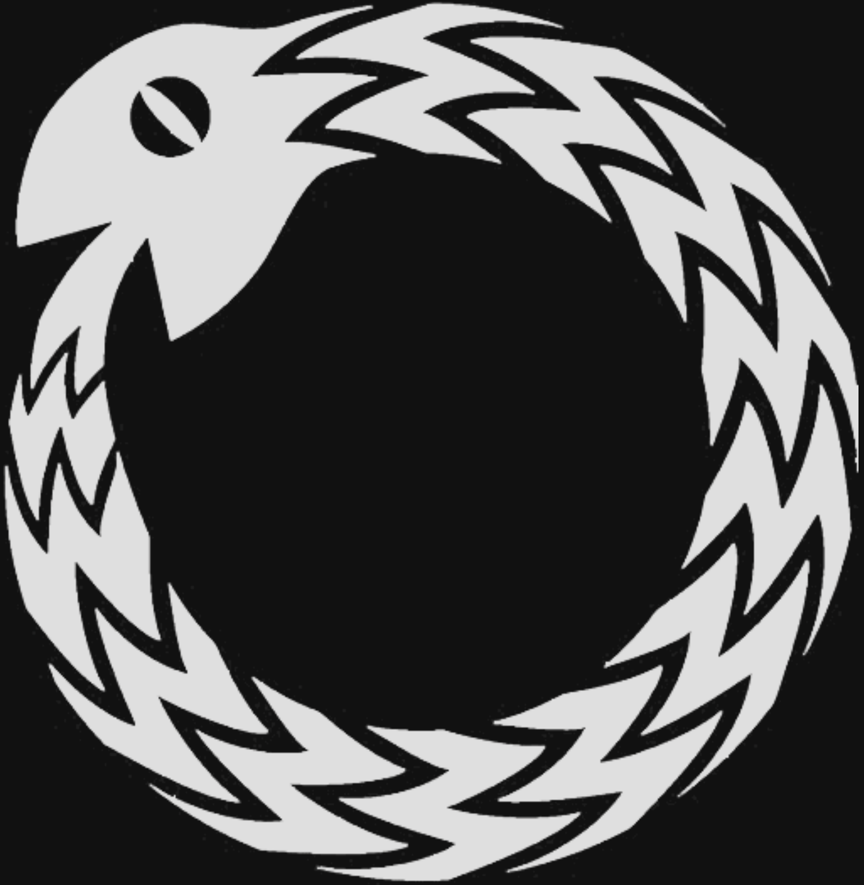
Low-resource language

Why don't we adapt resources from Sanskrit?

Top Resources

Dictionaries

Morphological analyzers



Credit: http://iflizwerequeen.com

Lingua Franca

Lingua Franca

Written in different scripts

Lingua Franca

Written in different scripts

Introduces variation!

Scripts

- Sinhalese
- Devanagari
- Burmese
- Transliterations
- ...

Literature

Literature

Scarce and not exhaustive

No annotated corpus

Generation

Generation

and Overgeneration

Irregular

Dictionary lookup

Regular

Dictionary lookup

Rule based generation:

- Lemma => Stem
- Stem + Ending => Form

Word class specific lemma ending

Lemma - Ending → Stem

Stem + Ending → Surface Form

Ending

Ending

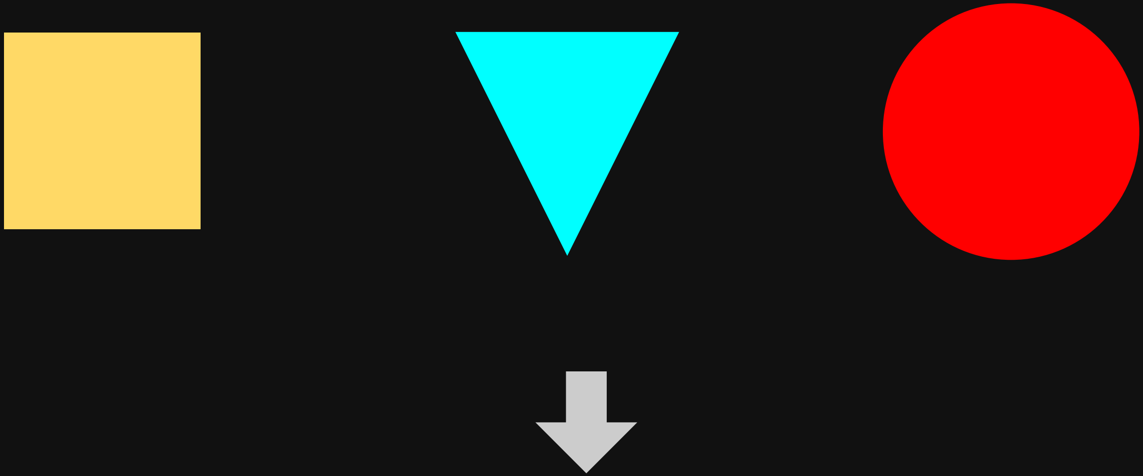
Ending

Stem + Ending → Form

Ending

Ending

Ending



Compiled Morphological Information

```xml
<paradigms>
    <paradigm type="noun">
        <number type="singular">
            <declension type="a">
                <gender type="masculine">
                    <case type="nominative">
                        <ending>o</ending>
                        <ending type="Drare">e</ending>
                    </case>
                    <case type="vocative">
                        <ending>a</ending>
                        <ending>ā</ending>
                        <ending type="Drare">e</ending>
                        <ending type="Drare">o</ending>
                    </case>
                    <case type="accusative">
                        <ending>aṃ</ending>
                    </case>
```

```xml
<paradigms>
    <paradigm type="noun">
        <number type="singular">
            <declension type="a">
                <gender type="masculine">
                    <case type="nominative">
                        <ending>o</ending>
                        <ending type="Drare">e</ending>
                    </case>
                    <case type="vocative">
                        <ending>a</ending>
                        <ending>ā</ending>
                        <ending type="Drare">e</ending>
                        <ending type="Drare">o</ending>
                    </case>
                    <case type="accusative">
                        <ending>aṃ</ending>
                    </case>
```

```
<paradigms>
    <paradigm type="noun">
        <number type="singular">
            <declension type="a">
                <gender type="masculine">
                    <case type="nominative">
                        <ending>o</ending>
                        <ending type="Drare">e</ending>
                    </case>
                    <case type="vocative">
                        <ending>a</ending>
                        <ending>ā</ending>
                        <ending type="Drare">e</ending>
                        <ending type="Drare">o</ending>
                    </case>
                    <case type="accusative">
                        <ending>aṃ</ending>
                    </case>
```

```
<paradigms>
    <paradigm type="noun">
        <number type="singular">
            <declension type="a">
                <gender type="masculine">
                    <case type="nominative">
                        <ending>o</ending>
                        <ending type="Drare">e</ending>
                    </case>
                    <case type="vocative">
                        <ending>a</ending>
                        <ending>ā</ending>
                        <ending type="Drare">e</ending>
                        <ending type="Drare">o</ending>
                    </case>
                    <case type="accusative">
                        <ending>aṃ</ending>
                    </case>
```

Lemma => Stem

Stem + Ending => Form

deva => dev-

dev- + -o => devo

Lemma => Stem

Stem + Ending => Form

deva => dev-

dev- + -o => devo

```
<declension type="ant">
    <gender type="masculine">
        <case type="nominative">
            <ending>aṃ</ending>
            <ending>ā</ending>
            <ending type="Cm2">anto</ending>
            <ending type="Drare">o</ending>
            <ending>ato</ending>
        </case>
```

karo + mi = karomi

I make

paca + mi = pacāmi

I cook

bhavaṃ (sir)

stem: bhav-

ending: -anto

form: ~~bhavanto~~

bhanto

Lemma

- Derive stem
- Select paradigm(s) based on word class
- Combine stem and endings
- Return generated forms and associated information

Verbs

Of Roots and Bases

Abstract Root

$\sqrt{kar}$ (to make)

Base

$\sqrt{kar} \to karo$	(to make)
$\sqrt{pac} \to paca$	(to cook)
$\sqrt{yudh} \to yujjha$	(to fight)

Seven declension classes

1+ bases

$\sqrt{cur}$ (to steal)

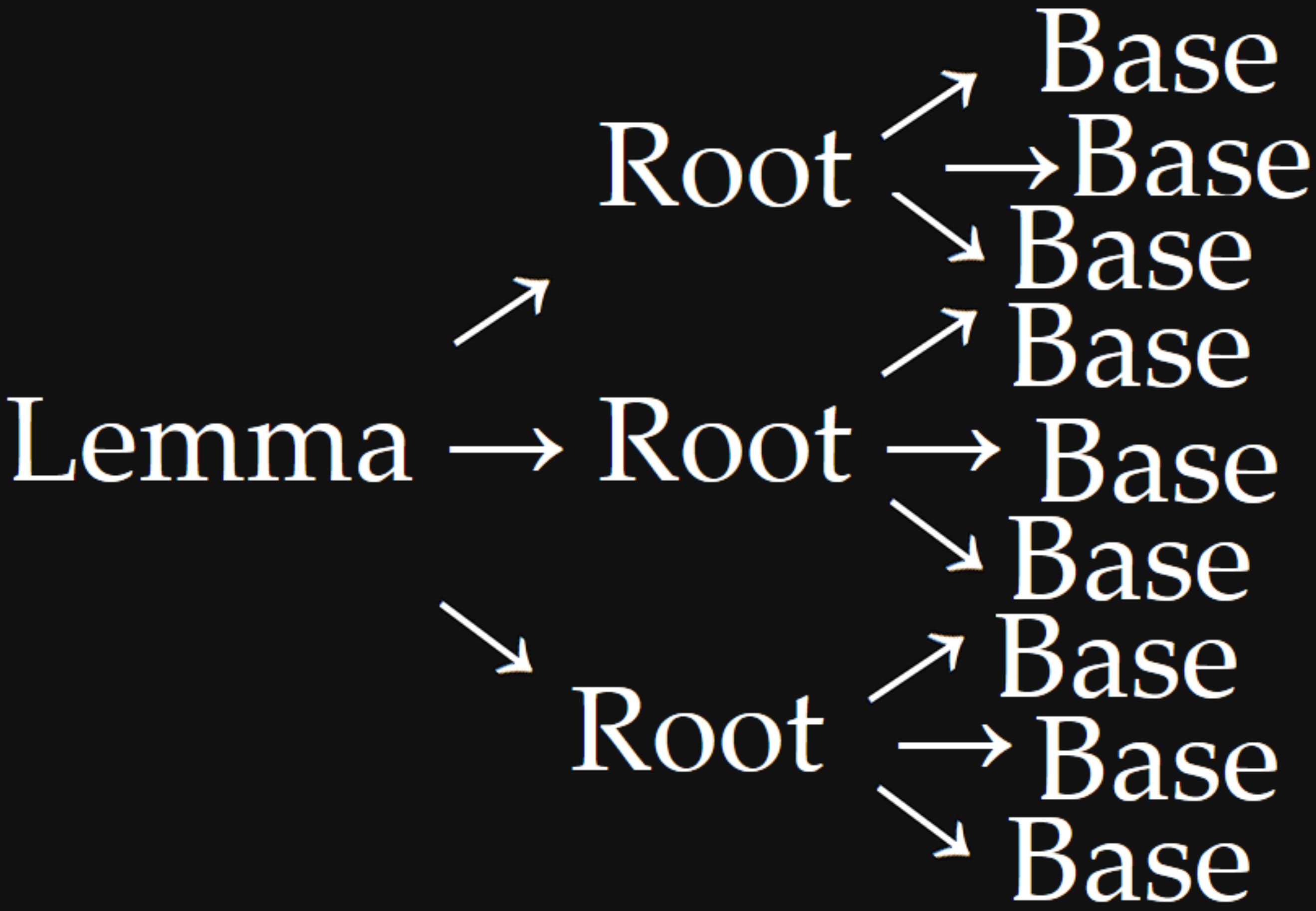
core-, coraya-

1+ bases

$\sqrt{rudh}$ (to obstruct)

rundha-, rundhi-, rundhī-, rundhe-, rundho-

Verb forms based on Root or Base?



Lemma

Root

Base

Base

Base

Root

Base

Base

Base

Root

Base

Base

Base

Irregular forms

Dictionary lookup

Full/Partial Irregularity

Output

JSON / XML

Key:Value pairs

Receiver can decide what information to use

```
{" lemma":"eka","forms ":{"numeral":[{
"gender ":"masculine", "number ":" singular",
"word ":" eko", "case":" nominative"},
{"gender ":"masculine", "number ":"
singular","word ":"ekassa", "case":" genitive"},...
```

Analysis

Lookup

Dictionary / Table lookup

Heuristic approach

Identify paradigmatic ending
→ Morphological Analysis
→ Separation Stem-Ending

```
<gender type="masculine">
    <case type="nominative">
        <ending>o</ending>
        <ending type="Drare">e</ending>
    </case>
    <case type="vocative">
        <ending>a</ending>
        <ending>ā</ending>
        <ending type="Drare">e</ending>
        <ending type="Drare">o</ending>
    </case>
    <case type="accusative">
        <ending>aṃ</ending>
    </case>
```

buddhe

```
<gender type="masculine">
    <case type="nominative">
        <ending>o</ending>
        <ending type="Drare">e</ending>
    </case>
    <case type="vocative">
        <ending>a</ending>
        <ending>ā</ending>
        <ending type="Drare">e</ending>
        <ending type="Drare">o</ending>
    </case>
    <case type="accusative">
        <ending>aṃ</ending>
    </case>
```

buddhe

Word Class Guesser

Heuristic Approach

Lemma

- Identify possible endings

Free Form

- Identify possible endings
- Weigh by length
- Weigh by frequency
- Prune results

Word Class Guesser: Lemma

Code Excerpt

```
if (ends(lemma, "a", "ā", "i", "ī", "u", "ū", "ant", "vā", "mā", "at"))
    guesses.add("adjective");
}
if (ends(lemma, "a", "i", "aṃ", "ma", "ya")) {
    guesses.add("numeral");
}
if (ends(lemma, "uṃ")) {
    guesses.add("indeclinable");
}
```

Results

	Accuracy
Nouns-Adjectives	99.96%
Pronouns	88.57%
Numerals	76.62%
Verbs	63.37%

Sandhi

Compound Sandhi

Intuition

- Identify possible sandhi loci
- Split into n words such that

$$\forall n : w_n \in D$$

Problems

- Requires extensive Dictionary
- More than one analysis possible
- Not a compound

External Sandhi

Corpus-based resolution

Sandhi-inducing words

- ca (and)
- hi (because)
- pi (also)

Hand-written rules

Regular Expressions

Replacement rules	
\bpañca\b	X
ñca\b	ṃ ca
X	pañca
ñhi\b	ṃ hi
ñpi\b	ṃ pi

Replacement rules	
\bpañca\b	X
ñca\b	ṃ ca
X	pañca
ñhi\b	ṃ hi
ñpi\b	ṃ pi

Internal Sandhi

~~Internal Sandhi~~

Conclusion

Paradigms for Generation and Analysis

Dictionary Integration for additional information

Rule-based and heuristic backup

RegEx-based External Sandhi Resolution

Lookup

Server Architecture

Well documented REST API

Easy integration

Data Processing

Extract structured data from unstructured data [n. ag. fr. abhijjhita in med. function] one who covets M <smallcaps>i.</smallcaps> 287 (T. abhijjhātar, v. l. °itar) = A <smallcaps>v.</smallcaps> 265 (T. °itar, v. l. °ātar).

[n. ag. fr. abhijjhita in med. function] one who covets M <smallcaps>i.</smallcaps> 287 (T. abhijjhātar, v. l. °itar) = A <smallcaps>v.</smallcaps> 265 (T. °itar, v. l. °ātar).

Pacati， [Ved． pacati， Idg． *peqŭō， Av． pac-； Obulg． peka to fry， roast， Lith， kepū bake， Gr． pέssw cook， pέpwn ripe] to cook， boil， roast Vin． IV， 264； fig． torment in purgatory （trs． and intrs． ）： Niraye pacitvā after roasting in N． S． II， 225， PvA． 10， 14． -- ppr． pacanto tormenting， Gen． pacato （+Caus． pācayato） D． I， 52 （expld at DA． I， 159， where read pacato for paccato， by pare daṇḍena pīḷentassa）． -- pp． pakka （q． v． ）． ‹-› Caus． pacāpeti & pāceti （q． v． ）． -- Pass． paccati to be roasted or tormented （q． v． ）． （Page 382）

Manual annotation

Open Problems

Verbs

Use verb form table

Attested forms only

Internal Sandhi

Illustrating Calculation

Splitting Internal Sandhi

"When two vowels meet, one may be elided."

When two vowels meet:

- elide first vowel
- elide second vowel
- no elision

8 vowels

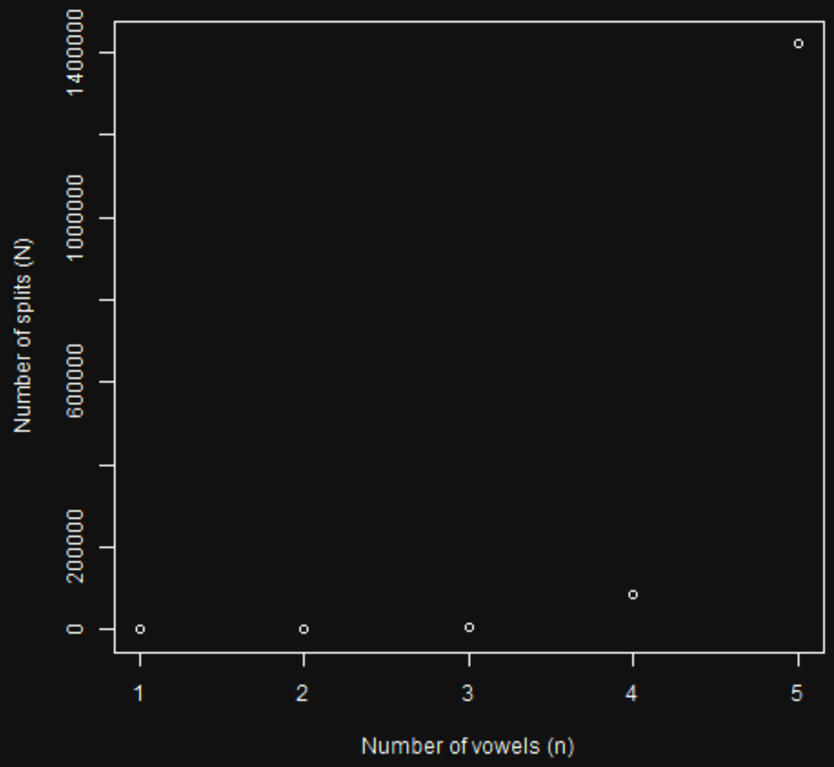
n-vowel-word

$$N = (1 + (2 * 8))^n$$

$$n = 1 \rightarrow N = 17$$

$$n = 2 \rightarrow N = 289$$

$$n = 3 \rightarrow N = 4913$$


Number of splits (N)
0
200000
600000
1000000
1400000
1
2
3
4
5
Number of vowels (n)

"A final dental is assimilated to the following consonant"

"A final dental is assimilated to the following consonant"

(DENTAL) (CONSONANT) : duplicate($2)

- kk: t k
- kk: th k
- kk: d k
- kk: dh k
- kk: n k
- kk: l k
- kk: s k
- ...

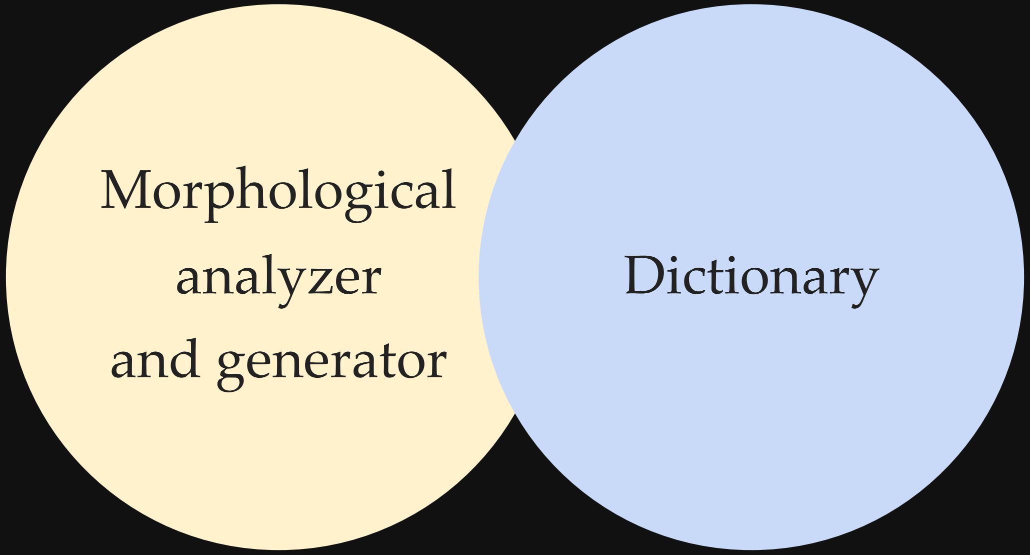
224 possibilities

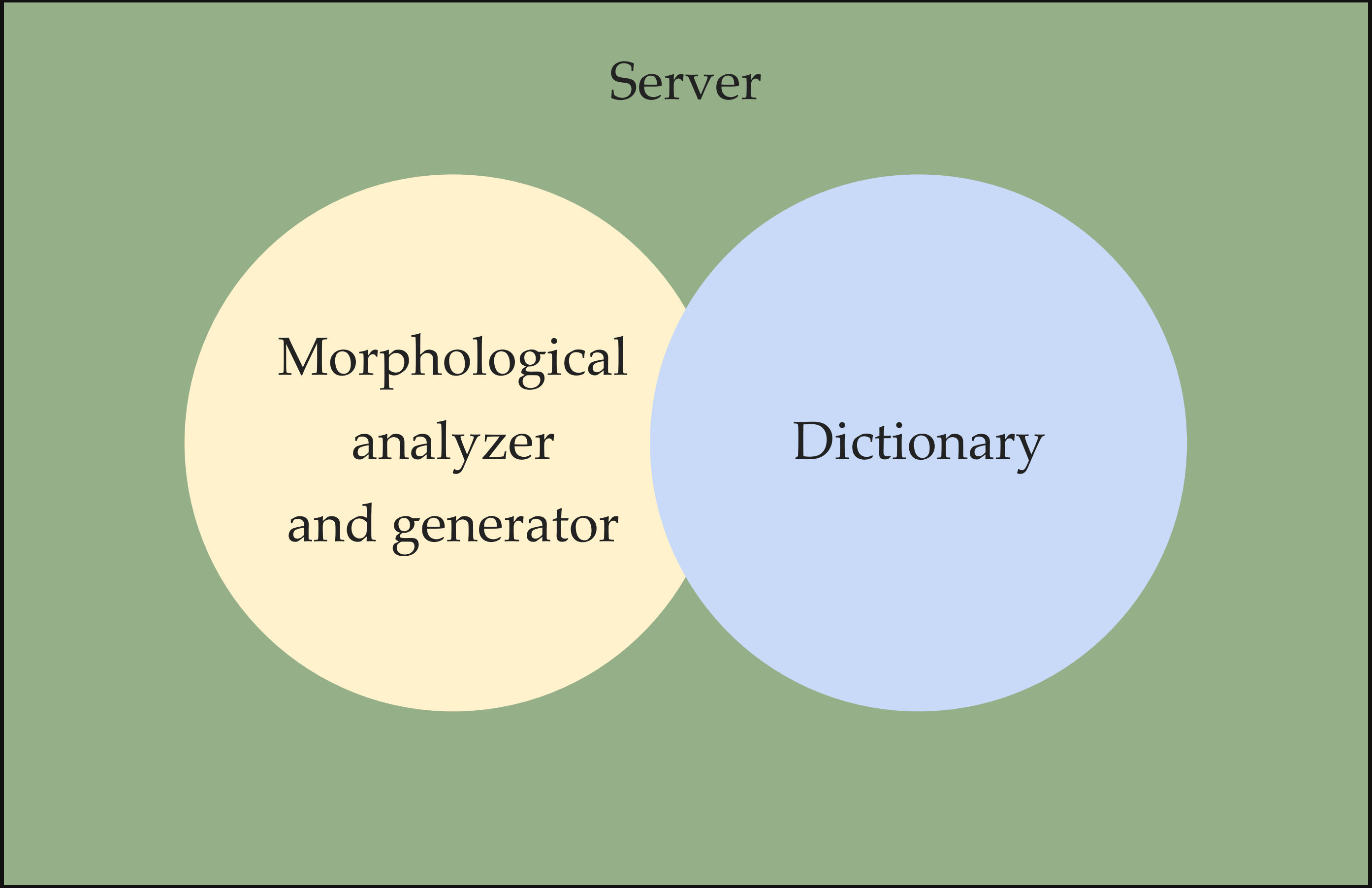
Sandhi merge rules

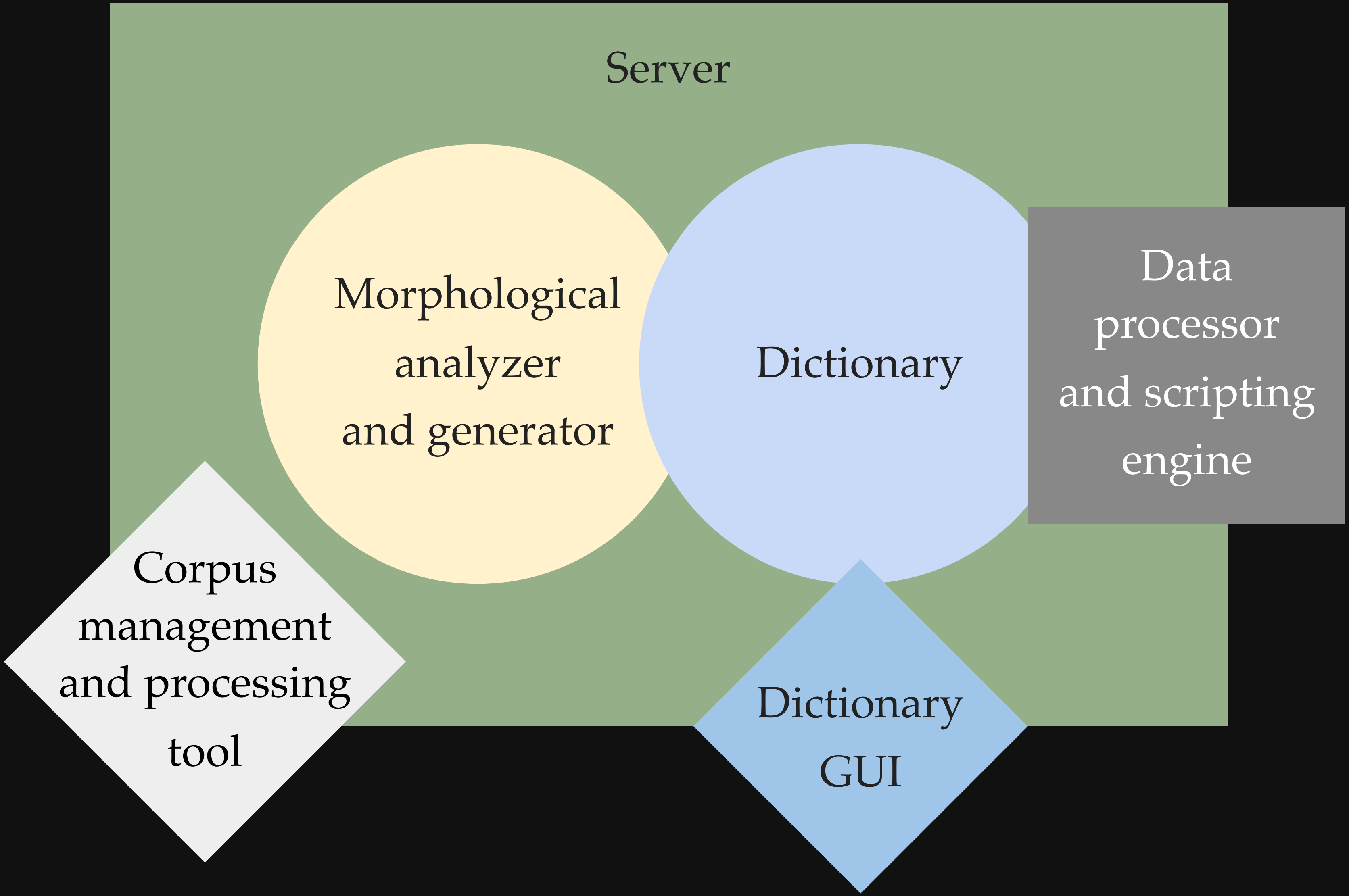
151 rules

Sandhi merge rules	Sandhi split rules
151 rules	1103 rules

Overall architecture


Morphological analyzer and generator
Dictionary


Server
Morphological analyzer and generator
Dictionary


Server
Morphological analyzer and generator
Dictionary
Data processor and scripting engine
Corpus management and processing tool
Dictionary GUI

Thank you for your attention!

Thank you for your attention!

Questions?