



# Sanskrit Knowledge-based Systems

## Annotation and Computational Tools

---

Hrishikesh Rajesh Terdalkar

**Supervisor:** Prof. Arnab Bhattacharya

November 21, 2023

Department of Computer Science and Engineering,  
Indian Institute of Technology Kanpur

# Introduction

---

Who was the father of Arjuna

Google search results for "Who was the father of Arjuna". The search bar shows the query. Below the search bar, it says "About 22,90,000 results (0.65 seconds)". The first result is "Arjuna / Fathers" with the text "king Pandu". Below this, there is a paragraph about Arjuna: "Arjuna is one of the heroes of the massive Indian epic named 'The Mahabharata', the longest Indian epic. He is the third of the five Pandava brothers, officially the son of king Pandu and his two wives Kunti (who is also known as Pritha) and Madri. Sep 15, 2013". Below the paragraph is a link to "Arjuna - Ancient History Encyclopedia" with the URL "https://www.ancient.eu / Arjuna".

अर्जुनस्य पिता कः

Google search results for "अर्जुनस्य पिता कः". The search bar shows the query. Below the search bar, it says "About 3,740 results (0.74 seconds)". The first result is "SANSKRIT: अर्जुनस्य दश नामानि (Ten Names of Arjun)" with the URL "iksusara.blogspot.com / 2015/02 / ten-names-of-arjun". Below this, there is a paragraph about Dussehra: "Dussehra 2019 On Dussehra How To Worship Shami Puja ...". Below the paragraph is a link to "अर्जुन - Sanskrit-Hindi Dictionary - Glosbe" with the URL "https://glosbe.com / Dictionary Sanskrit / Sanskrit-Hindi Dictionary".

September 2019

## Who was the father of Arjuna


Google

Who was the father of Arjuna

About 2,82,00,000 results (0.67 seconds)

Arjuna / Father

**Pandu**

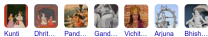


Arjuna, one of the five Pandava brothers, who are the heroes of the Indian epic the Mahabharata. Arjuna, son of the god Indra, is famous for his archery (he can shoot with either hand) and for the magical weapons that he wins from the god Shiva.

<https://www.britannica.com> ... Folk Literature & Fable

[Arjuna | Hindu mythology - Encyclopedia Britannica](#)

People also search for



Feedback

## अर्जुनस्य पिता कः

Google

अर्जुनस्य पिता कः

About 3,410 results (0.48 seconds)

अनुवाद-वीडियो: अर्जुनस्य पुत्रः भारीतः 30-Aug-2018

<https://www.upboardsolutions.com> > class-8-sanskrit-chap...  
**UP Board Solutions for Class 8 Sanskrit Chapter 13 वीरोधमिमन्तुः**

<https://www.gitasupersite.lit.ac.in> > s... > Translate this page  
**मूल श्लोकः - श्रीमद् भागवद्गीता | Gita Supersite**  
एवमपि विष्णुः नृपुण्यं कुरुष्व इति वचनं श्रुत्वा ... येन येन योगेन अर्जुनस्य अयमवस्था ...

<https://www.gitasupersite.lit.ac.in> > s... > Translate this page  
**Sanskrit Commentary By Sri - श्रीमद् भागवद्गीता | Gita Supersite**  
... को (अनेन विष्णुः के बड़े भाई होने से) विष्णु के ... अर्जुनस्य नृपुण्यं कुरुष्व इति वचनं श्रुत्वा ...

<https://hi.krishnakosh.org> > शृणु - Translate this page  
**अर्जुनस्य सखा - Krishnakosh**

December 2022

## Who was the father of Arjuna

Google search results for "Who was the father of Arjuna".

Search query: Who was the father of Arjuna

Results: About 2,67,00,000 results (0.39 seconds)

Arjuna was the son of Kunti, the wife of Kuru King Pandu, and the god Indra, who fathered him due to Pandu's curse.

Wikipedia  
https://en.wikipedia.org/wiki/Arjuna

Arjuna - Wikipedia

People also ask

- Who is the real father of Bhima?
- Who was father of Karna?

## अर्जुनस्य पिता कः

Google search results for "अर्जुनस्य पिता कः".

Search query: अर्जुनस्य पिता कः

Results: About 4,480 results (0.29 seconds)

Blogger  
http://newsdata.blogspot.com/ten-n... Translate this page

अर्जुनस्य दश नामानि (Ten Names of Arjuna)  
21-Feb-2015 — दशानं नाम कृष्णं रक्षा गतं मेरुं पिता जी के द्वारा कही कि हमारा हाथ से ही भगवान् कृष्ण से जुड़ा था इस लिङ्क

https://www.youtube.com/watch?v=... Translate this page

Arjun ke pita ka naam kya tha - YouTube  
Hello everyone I am ashish and you are watching meaning of names .  
Topic cover in this video > 1 - Arjun ...  
25-Apr-2021 - Uploaded by Meaning of names

Quora  
https://hi.quora.com/पिता-के-नाम-क्या-...  
"पिता" से संबंधित संस्कृत के श्लोक क्या हैं?  
30-Aug-2020 — "पिता" से संबंधित संस्कृत के श्लोक क्या हैं? अर्जुनस्य पिता के दश नाम (दशानम) ज्ञात रहति ...  
8 answers · Top answer: श्लोक अनेक हैं, इस विषय को देखो ...

November 2023

Query: Sanskrit    Response: ~Hindi

November 2023



You

अर्जुनस्य पिता कः

= *Who was Arjuna's father?*



ChatGPT

अर्जुनस्य पिता का नाम राजा पाण्डु था।

= *Arjuna's father's name was King Pāṇḍu*

Acceptable answer. 🧐

So, okay!? 🤔



You

कस्य पिता अर्जुनः

= *Whose father was Arjuna?*



ChatGPT

अर्जुनस्य पिता के रूप में राजा पाण्डु थे।

= *Arjuna's father was Pāṇḍu*



You

अर्जुनः कस्य पिता

= *Arjuna was whose father?*



ChatGPT

अर्जुनः के पिता का नाम राजा पाण्डु था।

= *Arjuna's father was Pāṇḍu*

Same question.

Alternate formulation!

Incorrect! 🤔

No 'understanding'.

# Information and Knowledge

## Information Retrieval (IR)

- Search text
- Ranked list of documents
- Indexes, Language models, ...
- Google, DuckDuckGo, ...

## Question Answering (QA)

- ‘Understand’, ‘search’, formulate
- Relevant phrases, sentences
- IR++, **Knowledge-bases**, LLMs, ...
- ChatGPT, Google Bard, ...

*“We are drowning in information but starved for knowledge.”*

*– John Naisbitt, Megatrends*

## Knowledge-bases (KB)

Structured storage of real-world information

- Use KBs to solve high-level problems: QA, Inference Engines, ...
- Inadequate performance for Indian languages
- Sanskrit
  - Vast and varied literature
  - Morphologically and semantically rich
  - Low-resource language (computational datasets)
- *Why not just use translations?*
  - Limited availability
  - Fail to convey the exact meaning
  - Questionable accuracy, misinterpretations
  - Performance of state-of-the-art tools



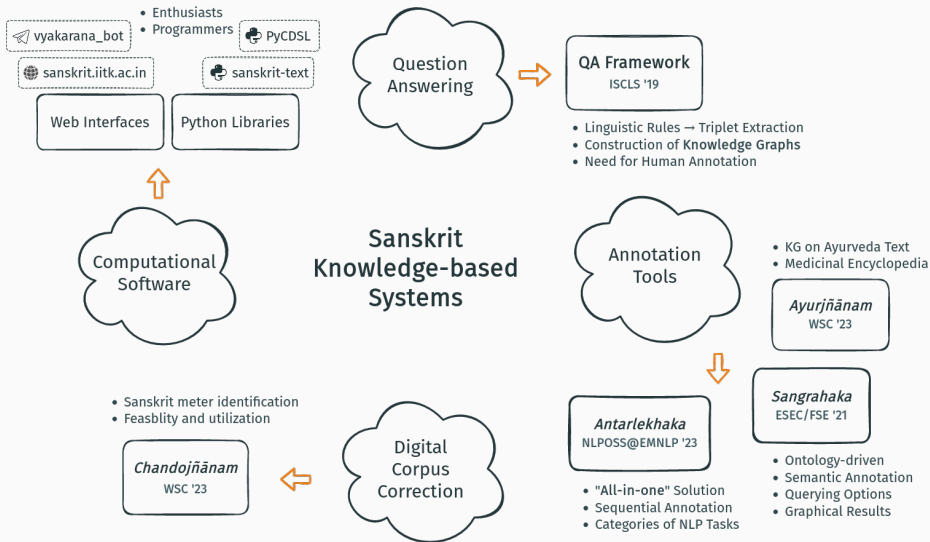
# Limitation of Translations: Anecdotal Evidence

## Entities detected by spaCy v3.4.4

*Ugrasrava/ORG, the son of Lomaharshana/GPE, surnamed Sauti/PERSON, well-versed in the Puranas/PERSON, bending with humility, one day/DATE approached the great sages of rigid vows, sitting at their ease, who had attended the twelve years'/DATE sacrifice of Saunaka/GPE, surnamed Kulapati/GPE, in the forest of Naimisha/GPE.*

– Mahābhārata

# Contributions



# Sanskrit Question-Answering

Automatic Construction of Knowledge Graphs

---

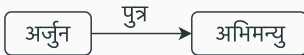
# Question-Answering using Knowledge Graphs

## Knowledge Graphs (KGs)

- Knowledge-bases with a graph data structure
- Real-world entities as nodes
- Relationships among the entities as directed edges

## Triplets (*subject, predicate, object*)

- Common way of encoding the relationship information
- Represents a directed edge
- e.g. (**Arjuna**, has-son, **Abhimanyu**)



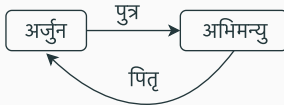
## Problem

Automatic extraction of triplets for construction of KGs

# Relationships

- Domain-specific, Application-specific (e.g., Kinship relations)
- Relationship words often corpus independent
  - पितृ (pitṛ, father), मातृ (mātr, mother), पुत्र (putra, son), etc.
- Multiple synonyms to the relationship words
  - पुत्री (putrī, daughter): दुहितृ (duhitṛ), तनया (tanayā), ...
- Implied Relationships

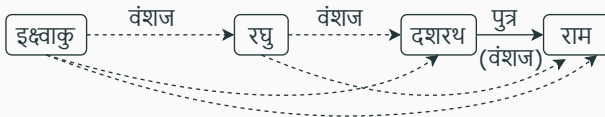
- Inverses

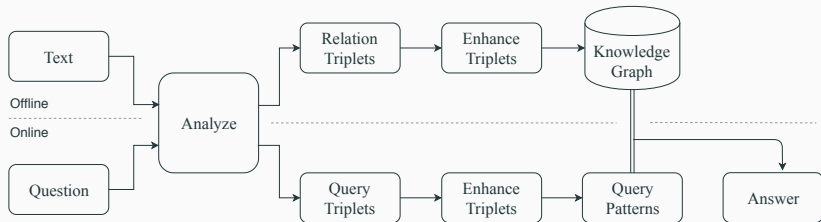


- Compositions



- Recursions

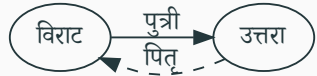




**Figure 1:** Overall framework of the Sanskrit QA system

# Example - Building Knowledge Graph

- Line from **Mahābhārata**  
विराटस्य दुहितरमुत्तरां नामाभिमन्युरुपेयेमे। (MBh 1.63.82a)
- **Sandhi-samāsa** splitting  
विराटस्य दुहितरम्-उत्तराम् नाम-अभिमन्युः-उपेय-इमे
- Morphological analysis  
विराट {g. sg. m.}, दुहितृ {acc. sg. f.}, उत्तरा {acc. sg. f.}
- Relationship Triplet  
(‘विराट’, ‘पुत्री’, ‘उत्तरा’)
- Inverse relationship  
‘पुत्री’ → [‘मातृ’, ‘पितृ’]
- Enhanced Triplet  
(‘उत्तरा’, ‘पितृ’, ‘विराट’)



# Example - Querying

- Lines from two different chapters

पूरोर्भार्या कौसल्या बभूव।

(MBh 1.63.8c)

शर्मिष्ठायाः सुतो द्रुह्युस्ततोऽनुः पूरुरेव च॥

(MBh 1.79.21b)

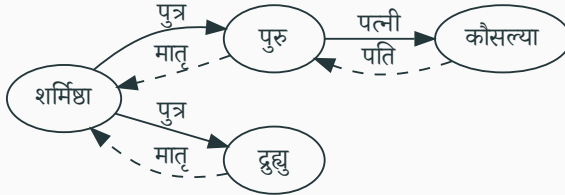
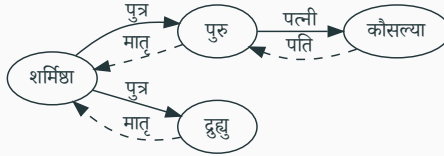


Figure 2: Knowledge graph enhanced with inverse relationships



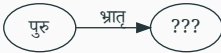
# Example - Querying



Q1: पुरोः भ्राता कः

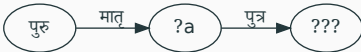
(Who was the brother of **Puru**?)

Triplet: ('पुरु', 'भ्रातृ', 'किम्')



Composition rules:

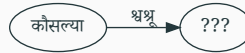
'भ्रातृ' → [('मातृ', 'पुत्र'), ('पितृ', 'पुत्र'), ...]



Q2: कौसल्यायाः श्वश्रूः कः

(Who was the mother-in-law of **Kausalyā**?)

Triplet: ('कौसल्या', 'श्वश्रू', 'किम्')



Composition rules:

'श्वश्रू' → [('पति', 'मातृ'), ('पत्नी', 'मातृ')]



# Performance

Text	Task	Total	Found	Correct	Precision	Recall	F1
Rāmāyaṇa	QParse	35	33	27	0.82	0.77	0.79
	QCond	27	19	09	0.47	0.33	0.39
	QAll	35	20	10	0.50	0.29	0.37
Mahābhārata	QParse	45	45	41	0.91	0.91	0.91
	QCond	41	36	22	0.61	0.54	0.57
	QAll	45	40	23	0.58	0.51	0.54
Combined	QParse	80	78	68	0.87	0.85	<b>0.86</b>
	QCond	60	55	31	0.56	0.46	<b>0.50</b>
	QAll	80	60	33	0.55	0.41	<b>0.47</b>

**Table 1:** Performance of the question-answering tasks on 80 questions collected from 12 users. *QParse*: query parsing task, *QCond*: conditional QA task (modulo *QParse*), *QAll*: overall question answering task.

# Errors in Knowledge Graph and Question-Answering

- Errors in parsing the question
  - कर्णार्जुनयोः कः सम्बन्धः → (किम्, किम्, सम्बन्ध)
  - Due to unhandled pattern
  - Easy to resolve, *if found*
  - *Difficult to be exhaustive*
- Errors in answering
  - हनुमतः पिता कः → (हनुमत्, पितृ, किम्)
  - Answer triplet (मारुति, पितृ, पवन) exists
  - मारुति is another name of हनुमत्
  - Use of dictionaries, thesauri 'can' help
    - *Abundant homonyms*
    - *Corpus-dependent*
- Errors in triplet identification
  - Both false-positives and false-negatives
  - *No evaluation dataset*

# Errors in Corpus and Tools

- Errors in the text
  - [चन्द्रि का] चर्महन्त्री च पशुमेहनकारिका
  - चन्द्रि का → चन्द्रिका
  - *Error detection and correction of corpora*
- Errors in state-of-the-art tools
  - Morphological analysis
    - नन्दिनी → नन्दिन् {acc. du. n.}
    - Expected: नन्दिनी {nom. sg. f.}
  - Oversplitting **sandhi** and **samāsa**
    - कारवी → का रवी
- *Compounding of errors*
  - कारवी → का रवी → किम् {nom. sg. f.}, रवि {acc. du. m.}
  - Expected: कारवी {nom. sg. f.}

# Issues in Automatic Knowledge Graph Construction

- Multiple components
  - World knowledge
  - Linguistic tasks
  - Computational tools
- Every component of the task has its own error rate
- Lack of evaluation datasets
  - Task-specific evaluation
- *No one-size-fits-all solution*

Manual annotation is necessary for performant solutions.

# Annotation Tools

---

## *Why create new annotation tools?*

- User-friendly interfaces
- Distributed annotation
- Web-based deployment
- Ease of setup
- Access management
- Scalability
- Crash tolerance
- **Task specific annotation needs**

# Sangrahaka – Annotation and Querying

- Ontology-driven annotation of *entities* and *relationships*
- Querying of KG using natural language query templates

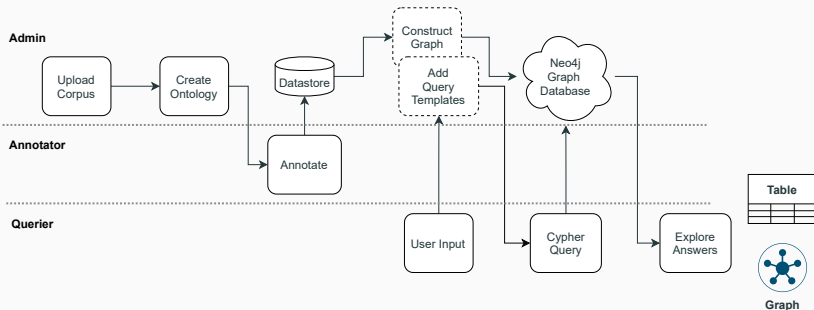


Figure 3: Architecture and Workflow of Sangrahaka



# Sangrahaka – Features and Capabilities

## Annotation

- Mark Nodes
- Mark Relations
- Adaptive Suggestions
- Multiple Annotators
- Multiple Curators

## Querying

- Template System
- Interactive Query
- Cypher Query
- Results Explorer
  - Graph
  - Table

## Administration

- Access Management
- Corpus Upload
- Ontology Creation
- Annotation Download

## Configuration

- Single File Configuration
  - SQL, SMTP, . . .
- Query Templates

## Customization

- Structured Code
- Modular
- Examples

## Other

- Unicode Support
- Web Deployment
- Fault Tolerance

Bhavaprakasha Nighantu - धान्यपर्वः

256343

	Line	Text				Split			?
४	256343	गोपुसः सुमनसि सान्निहितः स च कीर्तितः				गोपुसः सुमन-सनि सान्निहितः स च कीर्तितः			✓
Word	गोपुसः	सुमनः	अनि	सान्	निहितः	स	च	कीर्तितः	
Root	गोपुस	सुमन्	सनि	सान्	निहित्	ख्	च	कीर्तित	
Gender	m.	m.			m.				
Case	1	2			2				
Number	sg.	sg.			pl.				
Noun?	true	true	false	false	true	false	false	false	

Query

Show some details about #n#.

English

गोपुस

MATCH (x)-[relation]-(entity) WHERE entity.lemma =~ "गोपुस" RETURN \*

Submit

Entity Relation

Prepare

Line 256343

Source

Entity

Type None

Prepare

Entities

Confirm

गोपुस SUBSTANCE ✓

सुमन SUBSTANCE ✓



Entity Relation

Prepare

Line 256343

Source

Relation None

Detail

Target

Prepare

Relations

Confirm

(सुमन) -- [IS\_SYNONYM\_OF ()] --> (गोपुस) ✓

Query Result

id	r	p	r1	r2	s
गोपुस	IS_SYNONYM_OF	गोपुस	गोपुस	गोपुस	गोपुस
सुमन	IS_SYNONYM_OF	सुमन	सुमन	सुमन	सुमन

Showing 1 to 3 of 3 rows

**Figure 4:** Interfaces: Corpus Viewer, Entity Annotator, Relation Annotator, Query Interface, Graphical Result, Tabular Result



Concept	Words or Phrases
increases <b>bala</b> increase <b>vāta</b>	balya, balada, balāvaha, balaprada, balakara, balakṛt vātala, vātakṛt, vātakara, vātajanaka, vātajanānī, vātātikopana, vātaprakopaṇa, vātakopana, . . .
decreases <b>pitta</b>	pittaghna, pittapraṇāśana, pittapraśamana, pittahara, pit- taghnī, pittāpaha, pittajit, pittahṛt, pittavināśinī, . . .
decreases <b>vāta</b> and <b>pitta</b>	vātapittaghna, pittavātaghna, pittavātavibandhakṛt, vā- tapittahara, vātapittahṛt

**Table 2:** Semantic variations in Sanskrit – Examples from **Dhānyavarga**.

- Multiple ways of representing a single concept
- **Samāsa** for multiple increment or decrements at the same time
- Context-sensitive semantics (e.g. **-ghna**)

# Unnamed Entities

mudgo bahuvidhaḥ śyāmo haritaḥ pītakaṣṭhā.  
śveto raktaśca teṣāntu pūrvaḥ pūrvo laghuḥ smṛtaḥ. ||39||

- Entities referenced by their properties only, and not named at all
- Five colored variants of **mudga**, but *not named explicitly*
- Create *unnamed entities* (denoted by **x**-prefixed nodes)
- Unique identifier, e.g., **X1-256358**, **X2-256358**, ...
- Relations to describe the properties, e.g.,  
    śyāma ⊢ is (varṇa) Property of → **X1-256358**  
    harita ⊢ is (varṇa) Property of → **X2-256358**
- Word **teṣām** in second line refers to the five variants
- Relations between unnamed entities  
    **X1-256358** ⊢ is Better (in property laghu) than → **X2-256358**
- Anonymous nodes treated like any other node

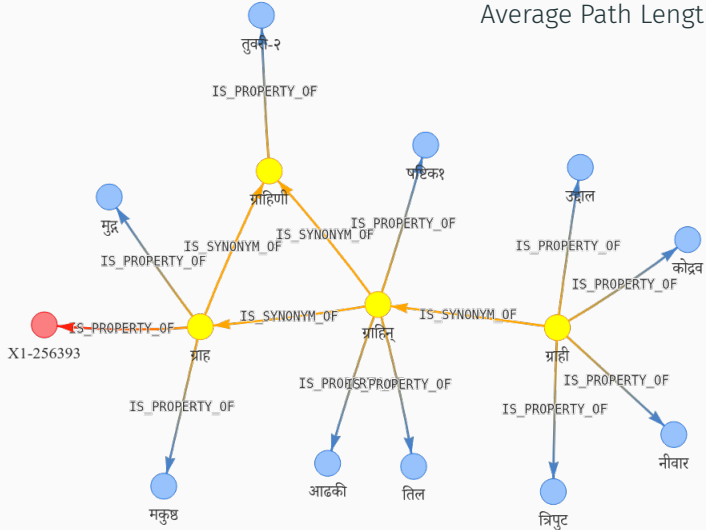
# Synonyms – Problem

- Relation ‘is **Synonym of**’ is symmetric and transitive
- A is a synonym of B  $\Leftrightarrow$  B is a synonym of A
- A is a synonym of B, B is a synonym of C  $\Leftrightarrow$  A is a synonym of C
- Several synonyms of each substance  
e.g.  $\text{rājikā} \leftrightarrow \text{kṣava} \leftrightarrow \text{kṣutābhijanaka} \leftrightarrow \text{kṛṣṇīkā} \leftrightarrow$   
 $\text{kṛṣṇasarṣapa} \leftrightarrow \text{rājī} \leftrightarrow \text{kṣujjanikā} \leftrightarrow \text{āsurī} \leftrightarrow \text{tīkṣṇagandhā} \leftrightarrow$   
 $\text{cīnāka}$
- Annotation:  $\text{uṣṇa} \vdash \text{is Property of} \rightarrow \text{rājikā}$
- Query: Find all properties of  $\text{cīnāka}$ .
- **Problem:**
  - Relations might be connected to each other only in a chain
  - Potentially 10 edge traversal required!

# Synonyms – Solution

- Identify connected components over 'is **Synonym of**'
  - Choose a canonical node (e.g. one with the highest out-degree)
  - Transfer all other edges from the group to the canonical node
- 
- Every node connected to canonical node.
  - Thus, at most 1 extra edge traversal required.
  - Initial computation cost for efficient querying.

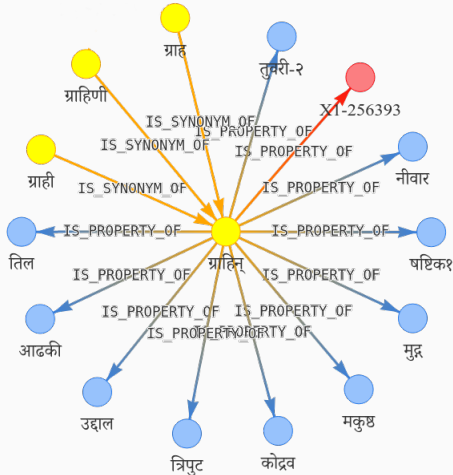
# Example – Before Optimization





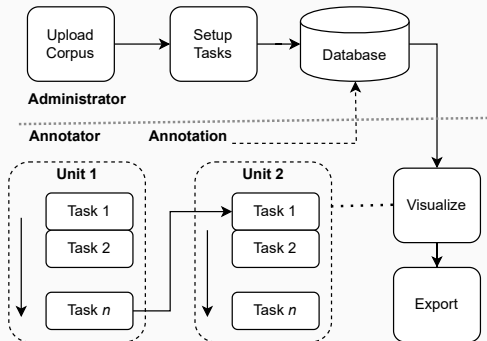
# Example – After Optimization

Average Path Length: 1.44



# Antarlekhaka – Multi-task Annotation

- Sequential annotation for text units (e.g., a verse)
- Multiple categories of NLP tasks
- Heuristics for aiding annotators



# Punctuation and Word Order

*"If no mistake you have made, losing you are.  
A different game you should play."*



[ na rocate mamāpyetadārye ] [ yadrāghavo vanam /  
tyaktvā rājyaśriyaṃ gacchet ] [ striyā vākyavaśaṃ gataḥ // 2  
... ]  
[ nāsyāparādham paśyāmi ] [ nāpi doṣaṃ tathāvidham ] /  
[ yena nirvāsyate rāṣṭrādvanaśāya rāghavaḥ ] // 4



[ ārye etad mama api na rocate ]  
[ yad rāghavo rājyaśriyaṃ tyaktvā vanam gacchet ]  
[ ... ]  
[ ... ]  
[ ... ]

**Figure 5:** Sanskrit verses from Valmiki Ramayana. Original text appears on the left with **sentence boundary** markers added. The **canonical word order** is shown on the right.

## Sentence Boundary Detection

## Canonical Word Ordering

### Token Annotation

- Lemmatization
- Morphological analysis
- Word segmentation

### Token Classification

- Named Entity Recognition
- Part-of-speech Tagging
- Compound Classification

### Token Graph

- Dependency Parsing
- Constituency Parsing
- Action Graph

### Token Connection

- Co-reference Resolution
- Interaction Networks

### Sentence Classification

- Sentiment Detection
- Sarcasm Detection

### Sentence Graph

- Discourse Graph
- Timeline Annotation

# Feature Comparison

	INCePTION	GATE	BRAT	FLAT	doccano	Sangrahaka	Antarlekhaka
Distributed Annotation	✓	✓	✓	✓	✓	✓	✓
Easy Installation			✓	✓	✓	✓	✓
Sequential Annotation						✓	✓
Querying Interface						✓	
Token Text Annotation	✓	✓	✓	✓			✓
Token Classification	✓	✓	✓	✓	✓		✓
Token Graph	✓	✓	✓	✓		✓	✓
Sentence Boundary	✓						✓
Canonical Word Order							✓
Sentence Classification	✓						✓
Sentence Graph							✓

**Table 3:** Comparison of NLP annotation tools based on primary features and supported tasks



# Antarlekhaka Annotation Interface

Antarlekhaka

HomeCorpusExportSettingsAdminLogout (Admin)

वाल्मीकिरामायणम् - Ayodhyā 18

Search



BoundaryAnvayaLemmaNERActionCoreferenceSentClassDiscourse

	Verse	Text	🔍
○	2488	तथा तु विलपन्ती तां कौसल्यां राममातरम् उवाच लक्ष्मणो दीनस् तत् कालसदृशं वचः	🟢
⦿	2489	न रोचते ममाप्य एतद् आर्यं यद् राघवो वनम् त्यक्त्वा राज्यश्रेयं गच्छेत् स्त्रिया वाक्यवशं गतः	🟢
Word	न	रोचते	ममाप्य मम
Lemma	न	रुच्	- मद्
UPOS	PART	VERB	- PRO

Sentence Boundary

2488

तथा तु विलपन्ती तां कौसल्यां राममातरम्  
उवाच लक्ष्मणो दीनस् तत् कालसदृशं वचः ##

2489

न रोचते ममाप्य एतद् आर्यं ## यद् राघवो वनम्  
त्यक्त्वा राज्यश्रेयं गच्छेत् ## स्त्रिया वाक्यवशं गतः

2490

विपरीतश् च वृद्धश् च विषयैश् च प्रथर्षितः  
नृपः किम् इव न ब्रूयाच् चौष्टमानः समन्मथः ##

2489

Submit



# Tools and technologies



Figure 7: Tools and technologies that go into *Sangrahaka* and *Antarlekhaka*

Both tools are live and being used in annotation tasks.



# Sanskrit Meter Identification

Utilization for text correction

---

# Chandojñānam: Meter Identification and Utilization

**Sanskrit Meters:** Binary signature<sup>3</sup> of every line

Majority of extant Sanskrit literature follows Sanskrit prosody

- Identify meters from Sanskrit text or images
- Catch errors in the text and suggest corrections!
- Web-based application, Python library
- Three input modes: (1) plain text, (2) images (3) text files
- Two OCR Engines: (1) Google Drive OCR (2) Tesseract OCR
- Transliteration support (powered by **indic-transliteration**)
- Two meter identification modes: (1) line mode (2) verse mode
- **Fuzzy matching support using edit-distance comparison**

<https://sanskrit.iitk.ac.in/jnanasangraha/chanda/>

<sup>3</sup>Every syllable is classified as **laghu** (short) or **guru** (long) based on its pronunciation

# Feature Comparison

Features		[Mis07]	[MSG13]	[Raj20]	[Nei22]	Chandojñānam
Availability	Web Interface	✓ <sup>4</sup>	✓ <sup>5</sup>	✓	✓	✓
	Software Library			✓	✓	✓
Input	Text	✓	✓	✓	✓	✓
	Arbitrary Lines					✓
	Multiple Verses					✓
	Textfile Upload				✓	✓
	Image Upload					✓
Functionality	Meter Identification	✓	✓	✓	✓	✓
	Error Tolerance			✓	✓	✓
	Fuzzy Matching			✓		✓

**Table 4:** Feature comparison of extant meter identification systems

<sup>4</sup><http://sanskrit.sai.uni-heidelberg.de/Chanda/HTML/> no longer functional.

<sup>5</sup><https://sanskritlibrary.org:8080/MeterIdentification/> no longer functional.

Finding approximate and close matches if no exact match

## Why?

- Digitally available Sanskrit text can be erroneous
  - Manual data entry
  - Post-scanning OCR followed by manual correction
- Types of Errors
  - Characters may be misspelt, e.g., रु (ru) as रू (rū)
  - Characters may be missing, e.g., वर्गे (vargai) as वगै (vagai)
  - Characters may be misidentified, e.g., ऋ (ṛ) as क्र (kra)
  - Characters may get split, e.g., ख (kha) as ख (rava)
- Several such errors can affect the metrical pattern of the text

# Chandojñānam Interface

<https://sanskrit.iitk.ac.in/jnanasangraha/chanda/>

Text

Output Scheme: Match Input




ध्यायेदाजानुबाहुं धृतडारधनुषं बद्धपद्मासनस्थं  
पीतं वासो वसानं नवकमलदलस्पर्धिनेत्रं प्रसन्नम् ।  
वामाङ्कारूढ सीतामुखकमरमिलोचनं नीरदाभं  
नानालङ्कारदीप्तं दधतमुरुजटामण्डनं रामचन्द्रम् ॥

☐ Verse Mode

☒ Line Mode

Identify

Results



Akṣarāṇi	ध्या	ये	दा	जा	नु	बा	हुं	धृ	त	डा	र	ध	नु	षं	ब	द्ध	प	द्वा	स	न	स्थं
Laghu-Guru	ग	ग	ग	ग	ल	ग	ग	ल	ल	ग	ल	ल	ल	ग	ग	ल	ग	ग	ल	ग	ग
Gaṇa	म				र			भ			भ			य			य			य	
Counts	21 अक्षराणि, 34 मात्राः																				
Jāti	प्रकृतिः																				
Chanda	स्रग्धरा (1 edit)																				

+ Fuzzy

# Feasibility for Text Correction

## Simulate digitization pipeline

- Generate PDF from Wikisource text
  - Run two OCR systems: Google, Tesseract
  - Obtain the OCR-ed versions of the text
- 
- Three versions<sup>6</sup> of **Meghadūta**<sup>7</sup> composed by **Kālidāsa**
    - Wikisource, sanskritdocuments.org and GRETIL
  - Texts with more metrical variety
    - Śāntavilāsa (36 verses) (12 distinct meters)
    - Śrīrāmarakṣāstotra (39 verses) (9 distinct meters)
    - Rājendrakarṇapūra (72 verses) (4 distinct meters)
  - Total 14 text versions, 1038 verses, exhibiting 17 distinct meters

---

<sup>6</sup>Single text from different sources can differ in several places

<sup>7</sup>Also used by [Raj20] for evaluation

# Results – Error Tolerance

		Meghadūta					Śāntavilāsa			Rāmarakṣā			Rājendrakarṇapūra			Total
		SD	GR	WS	GO	TO	WS	GO	TO	WS	GO	TO	WS	GO	TO	
Number of Verses		117	111	123	123	123	36	36	36	39	39	39	72	72	72	1038
Unique Chanda		1	1	1	1	1	12	12	12	9	9	9	4	4	4	17
Erroneous Verses		20	79	2	31	77	13	16	31	1	4	13	12	26	71	396
Correct	[Nei22]	20	79	2	30	66	11	13	14	0	2	9	12	24	36	318 (80.3%)
Meters	[Raj20]	19	79	2	30	75	12	15	24	1	2	9	12	26	58	364 (91.9%)
Identified	Chandojñānam	20	79	2	31	77	13	16	29	1	3	9	12	26	71	389 (98.2%)

**Table 5:** Error tolerance of meter identification systems. (Versions are WS: Wikisource, GO: Google OCR, TO: Tesseract OCR, SD: sanskritdocuments.org, GR: GRETIL.) **Chandojñānam** is able to detect correct **chanda** from erroneous verses 98.2% of the times.

## Examples of actual errors from Wikisource version of Meghadūta

### Error #1

- Line: कालक्षेपं ककुभसुरभौ पर्वते पर्वेते ते (Pāda 3, Śloka 1.23)
- Incorrect word पर्वेते (should be पर्वते)
- Likely due to OCR error and an oversight by the curator
- Suggestion: [[['का', 'ल', 'क्षे', 'पं'], ['क', 'कु', 'भ', 'सु', 'र', 'भौ'], ['प', 'र्व', 'ते'], ['प', 'र्वेते'], ['ते']]]
- Correctly points to the location where a change is required



## Error #2

- Line: साभिज्ञानप्रहितकुशलैस्तद्वचोभिर्ममापि (Pāda 3, Śloka 2.53)
- Extra letter (त) present in the **sandhi** of words कुशलैः and तद्वचोभिः
- Suggestion: [[[ 'सा', 'भि', 'ज्ञा', 'न', 'प्र', 'हि', 'त', 'कु', 'श', 'लै', 'd(स्त)', 'त', 'द्व', 'चो', 'भि', 'र्म', 'मा', 'पि' ]]]
- Points out correctly that a syllable needs to be deleted
- However, points to an incorrect syllable स्त to be deleted
  - Both स्त and त are **laghu** letters
  - Deletion of either letter  $\implies$  the correct metrical signature
  - Impossible for a meter identification based system

# Miscellaneous Tools

<https://sanskrit.iitk.ac.in/>

---

## Representation of numeric values using letters, syllables or words

- Ease of remembrance
- Many-to-one mapping of string to numbers
- Systems
  - **Kaṭapayādi Saṅkhyā**: alpha-syllabic system
  - **Āryabhaṭīya Saṅkhyā**: alpha-syllabic system
  - **Bhūtasāṅkhyā**: words with numeric connotation
- Interfaces
  - Decode: Numeric values from given strings (*deterministic*)
  - **Encode**: Generate strings from a numbers
    - System dependent, e.g., **Kaṭapayādi**: *Data-driven*

<https://sanskrit.iitk.ac.in/jnanasangraha/sankhya/>

# Kaṭapayādi – Example

Select corpora

सामाधनम्

महाभारतम्

भावप्रकाशनिघण्टुः

श्रीमद्भगवतम्

Preferred number of words

Small

Encode number

14111265

Submit

Kaṭapayādi Encodings

मां तारयत्यार्यः स्वल्पं

मचक्रुकस्य यं वापि

शचिष्ठया कार्यं त्वया

Decode text

मां तारयत्यार्यः स्वल्पं

Submit

Kaṭapayādi Number

14111265

Split	मा	ं	ता	र	य	त्	या	र्	य	ः	स्	व	ल्	प	ं
Relevant	म		त	र	य		य		य			व		प	
Numbers	5		6	2	1		1		1			4		1	

Figure 8: Kaṭapayādi System – Encoding and Decoding

# Vaiyyākaraṇaḥ: Sanskrit Grammar Bot for Telegram



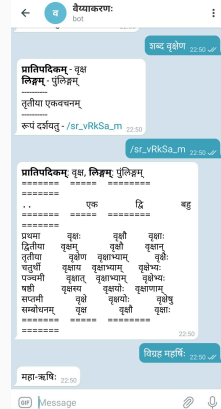
@vyakarana\_bot

[https://t.me/vyakarana\\_bot](https://t.me/vyakarana_bot)

- Telegram bot
- State-of-the-art Sanskrit tools

## Features

- Stem finder (Prātipadikam)
- Declension generator (Subantāḥ)
- Root finder (Dhātuḥ)
- Conjugation generator (Tīnantāḥ)
- Word segmentation (Sandhisamāsa)



# PyCDSL: Pythonic Access to Cologne Digital Sanskrit Lexicon

```
pip install PyCDSL
```

## Cologne Digital Sanskrit Lexicon (CDSL)

- Sanskrit-English, English-Sanskrit, Sanskrit-Sanskrit
- Specialized Dictionaries

## PyCDSL Features

- Download, manage, search
- Command Line Interface (CLI)
  - Console Command (`cdsl`)
  - REPL Interface (`cdsl -i`)
- Module to use in Python projects

```
import pycdsl

# default install at ~/cdsl_data
CDSL = pycdsl.CDSLCorpus()

# setup
CDSL.setup()

# dictionary accessible using `[]` operator
results = CDSL["MW"].search("राम")

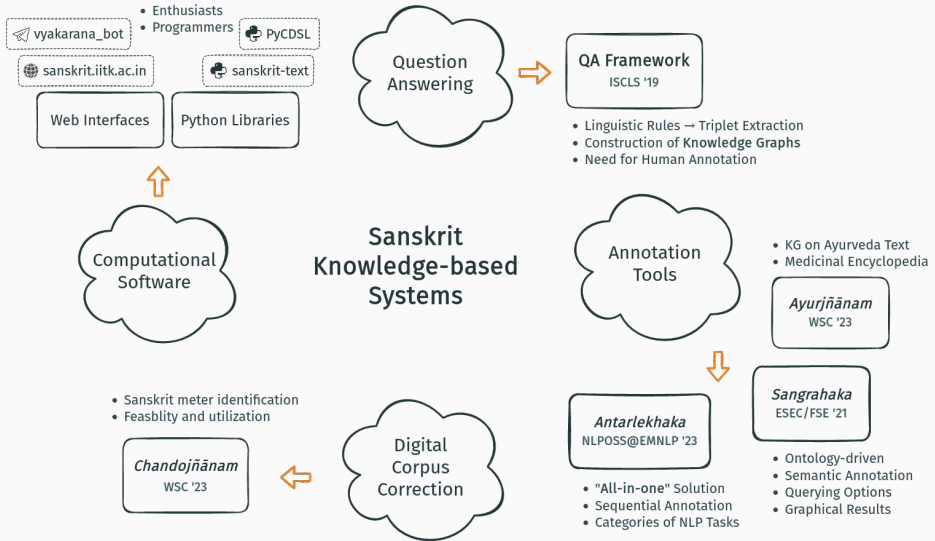
# alternatively, also accessible as attribute
results = CDSL.MW.search("राम")

# iterate over a `CDSLDict` instance
for entry in CDSL.MW:
    print(type(entry))
    print(entry)
    break
```

# Conclusions

---







# Summary





- High-level tasks for low-resource languages
  - Creation of computationally usable datasets
  - NLP tasks in Indian context
- Multilingual and Cross-Lingual NLP
  - Unified grammar for Indian languages
  - Machine translation among Indian languages
- Hybrid approaches
  - Knowledge-based Systems
  - Large Language Models (LLMs)

# Publications

-  **Hrshikesh Terdalkar and Arnab Bhattacharya.**  
**Framework for question-answering in Sanskrit through automated construction of knowledge graphs.**  
*In Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 97–116, IIT Kharagpur, India, October 2019. Association for Computational Linguistics.
-  **Hrshikesh Terdalkar and Arnab Bhattacharya.**  
**Sangrahaaka: A tool for annotating and querying knowledge graphs.**  
*In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*, page 1520–1524, New York, NY, USA, 2021. Association for Computing Machinery.
-  **Hrshikesh Terdalkar, Arnab Bhattacharya, Madhulika Dubey, S Ramamurthy, and Bhavna Naneria Singh.**  
**Semantic annotation and querying framework based on semi-structured Ayurvedic text.**  
*In Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 155–173, Canberra, Australia, January 2023. Association for Computational Linguistics.
-  **Jivnesh Sandhan, Ashish Gupta, Hrshikesh Terdalkar, Tushar Sandhan, Suwendu Samanta, Laxmidhar Behera, and Pawan Goyal.**  
**A novel multi-task learning approach for context-sensitive compound type identification in Sanskrit.**  
*In Proceedings of the 29th International Conference on Computational Linguistics*, pages 4071–4083, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
-  **Hrshikesh Terdalkar and Arnab Bhattacharya.**  
**Chandojnanam: A Sanskrit meter identification and utilization system.**  
*In Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 113–127, Canberra, Australia, January 2023. Association for Computational Linguistics.
-  **Hrshikesh Terdalkar and Arnab Bhattacharya.**  
**Antarlekhaka: A comprehensive tool for multi-task natural language annotation.**  
*In Proceedings of the 3rd Workshop on NLP Open Source Software at the 2023 Conference on Empirical Methods in Natural Language Processing, NLP-OSS @ EMNLP*, Singapore, December 2023. Association for Computational Linguistics.

Thank you!

Questions?

# Appendix I

## Automatic KG Construction

---

# Processing Sanskrit Text

- Sentence: कर्णार्जुनयोः कश्श्रेष्ठः (Who was greater – Karṇa or Arjuna?)
- Splitting of **samāsa** and **sandhi**
  - *Sanskrit Sandhi and Compound Splitter* <sup>8</sup>
  - Output: कर्ण-अर्जुनयोः कः-श्रेष्ठः
- Morphological analysis of the word
  - *The Sanskrit Heritage Platform* <sup>9</sup>
    - case (**vibhakti**, विभक्ति)
    - number (**vacana**, वचन)
    - gender (**liṅga**, लिङ्ग)
  - Output:
    - कर्ण {voc. sg. m.}
    - अर्जुन {loc. du. m.}
    - किम् {nom. sg. m.}
    - श्रेष्ठ {nom. sg. m.}

---

<sup>8</sup>Oliver Hellwig, Sebastian Nehrlich: *Sanskrit Word Segmentation Using Character-level RNN and CNNs*. EMNLP 2018.

<sup>9</sup>The Sanskrit Reader Companion, Heritage Platform, Gérard Huet, <https://sanskrit.inria.fr/DICO/reader.fr.html>

# Building Knowledge Graph

## World Knowledge

- List of kinship relationship words and synonyms
- Inference rules
  - Inverse relations
  - Composite relations

## Triplets Extraction

- Search for relationship words
- Proximity of subject and object (assumption)
  - Context window of  $n$  verses
- Case-based rules
  - **subject**: genitive case (ṣaṣṭhī vibhakti)
  - **predicate**: relationship word (various cases)
  - **object**: same case as the **predicate**

# Knowledge Graph Details

		Rāmāyaṇa	Mahābhārata
Time taken	Preprocessing	~ 3.5 days	~ 13 days
	Triplet Extraction	14.18 sec	57.19 sec
	Triplet Enhancement	0.40 sec	2.05 sec
Before enhancement	Entities (Nodes)	1,711	3,552
	Triplets (Edges)	6,155	18,936
	Distinct Relations	24	25
After enhancement	Entities (Nodes)	1,711	3,552
	Triplets (Edges)	11,367	32,395
	Distinct Relations	27	27

**Table 6:** Statistics of the knowledge graphs for the human relationships.

# Appendix II

## Sanskrit Meter Identification and Utilization

---



# Background

- Classification of syllables
  - Pronunciation dependent
  - **Laghu** (*short*)
    - Letters with short vowels
  - **Guru** (*long*)
    - Letters with long vowels
    - **Laghu** letters followed by a joint letter (**saṃyogaḥ**)
    - Last letter of a **pāda** (conditional)
- **Mātrā**: Laghu 1, Guru 2
- **Gaṇa**: Sequence of three letters ( $2^3 = 8$ )
- Chanda Types
  - **Akṣaracchanda**: Sequences of laghu-guru
    - **Samavṛtta**, **Ardhasamavṛtta** and **Viṣamavṛtta**
  - **Mātrācchanda**: Counts of mātrā
- Literature: **Vṛttaratnākaraḥ**, **Chandovicitīḥ**, **Chandomañjarī** etc.

# Fuzzy Matching

## How?

- **Problem:** Finding the *nearest matching string* for the *metrical signature* of the text line
- Compute Levenshtein edit-distance of the observed pattern
- Normalize the edit-distance by the length of target pattern

$$\text{Similarity} = 1 - \frac{\text{Levenshtein distance}}{\text{Length of target match}}$$

- Topmost  $k$  matches as the possible fuzzy matches
- Suggestions: changes to transform the input into the target
  - insert, delete, replace