



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

Google Search

Who was the father of Arjuna

Google search results for "Who was the father of Arjuna". The search bar shows the query. Below it, there are filters for All, News, Videos, Images, Maps, More, Settings, and Tools. It indicates about 22,900,000 results found in 0.65 seconds. The top result is from "Arjuna / Fathers" which lists "king Pandu". Below this, a snippet from "Arjuna - Ancient History Encyclopedia" provides a detailed description of Arjuna's life, mentioning he is the third of the five Pandava brothers, the son of king Pandu and his two wives Kunti (also known as Priti) and Madri. The date of Sep 15, 2013, is also mentioned. At the bottom, there is a link to the source and a feedback button.

अर्जुनस्य पिता कः

Google search results for "अर्जुनस्य पिता कः". The search bar shows the query. Below it, there are filters for All, Videos, News, Maps, Images, More, Settings, and Tools. It indicates about 3,740 results found in 0.74 seconds. The top result is a link to "SANSKRIT: अर्जुनस्य दश नामानि (Ten Names of Arjun)" from iksusara.blogspot.com. The snippet below it discusses the ten names of Arjun. Other results include links to "Dussehra 2019 On Dussehra How To Worship Shami Puja" and "अर्जुन - Sanskrit-Hindi Dictionary - Glosbe".

September 2019

Google Search

Who was the father of Arjuna

Who was the father of Arjuna

Q All News Images Videos Books More Tools

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

Arjuna J 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 | 24-Nov-2022

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

People also search for

Kunti Dhriti... Pandu... Gandhi... Vichitravirya... Arjuna Bhishma...

Feedback

अर्जुनस्य पिता कः

Google

অঙ্গুলস্য পিতা ক:

Q All Videos News Images Shopping | More Tools

About 3,410 results (0.48 seconds)

4 ৩ মুরি

অনুবাদ চীজ: অধিগব্ধ: অঙ্গুলস্য পুঁত: মারীদা। 30-Aug-2018

<https://www.upboardsolutions.com/class-8-sanskrit-chap...>

UP Board Solutions for Class 8 Sanskrit Chapter 13 বীরোচনিমন্ত্ৰ:

About | Featured snippets | Feedback

[https://www.gitasupersite.itt.ac.in/g... · Translate this page](https://www.gitasupersite.itt.ac.in/g...)

মূল শ্লোক - শীর্ষদ ভারবীরো | Gita Supersite

যান মান মিন তৃপ্তুমুনোগতি গু তুুু ... মে মে মোৰু অঙ্গুল আমাকা ...

[https://www.gitasupersite.itt.ac.in/g... · Translate this page](https://www.gitasupersite.itt.ac.in/g...)

Sanskrit Commentary By Sri - শীর্ষদ ভারবীরো | Gita Supersite

... গো (অজ্ঞ সিদ্ধ কে এই পার হোলো সো) জিবি থে ... অঙ্গুল পুতুলো বলো জিবি ...

[https://hi.krishnakosh.org/কৃষ্ণ · Translate this page](https://hi.krishnakosh.org/কৃষ্ণ)

অঙ্গুলস্য সখ্যা - Krishnakosh

December 2022

Google Search

Who was the father of Arjuna

Google search results for "Who was the father of Arjuna". The search bar shows the query. Below it are filters for All, Images, Videos, News, Books, and More, along with a Tools button. The result count is about 2,67,00,000 results in 0.39 seconds. The top result is a snippet from Wikipedia: "Arjuna was the son of Kunti, the wife of Kuru King Pandu, and the god Indra, who fathered him due to Pandu's curse." To the right of the text is a small image of a statue of Arjuna. Below the snippet is a link to "Arjuna - Wikipedia". Further down, under "People also ask", are questions like "Who is the real father of Bhima?" and "Who was father of Karna?".

अर्जुनस्य पिता कः

Google search results for "अर्जुनस्य पिता कः". The search bar shows the query in Sanskrit. Below it are filters for All, Videos, Images, News, Shopping, and More, along with a Tools button. The result count is about 4,480 results in 0.29 seconds. The top result is a blog post by "Blogger" titled "अर्जुनस्य दश नामानि (Ten Names of Arjun)". The post discusses the ten names of Arjuna. Below the post is a link to "Arjun ke pita ka naam kya tha - YouTube". A thumbnail for the video shows a person speaking. Below the video is a Quora answer from "Iqbal" asking "पिता से संबंधित संकृत के ल्लोक कवाह हैं ?". The answer is from 20-Aug-2020 and mentions Arjuna's names.

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 Pāṇḍū*

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

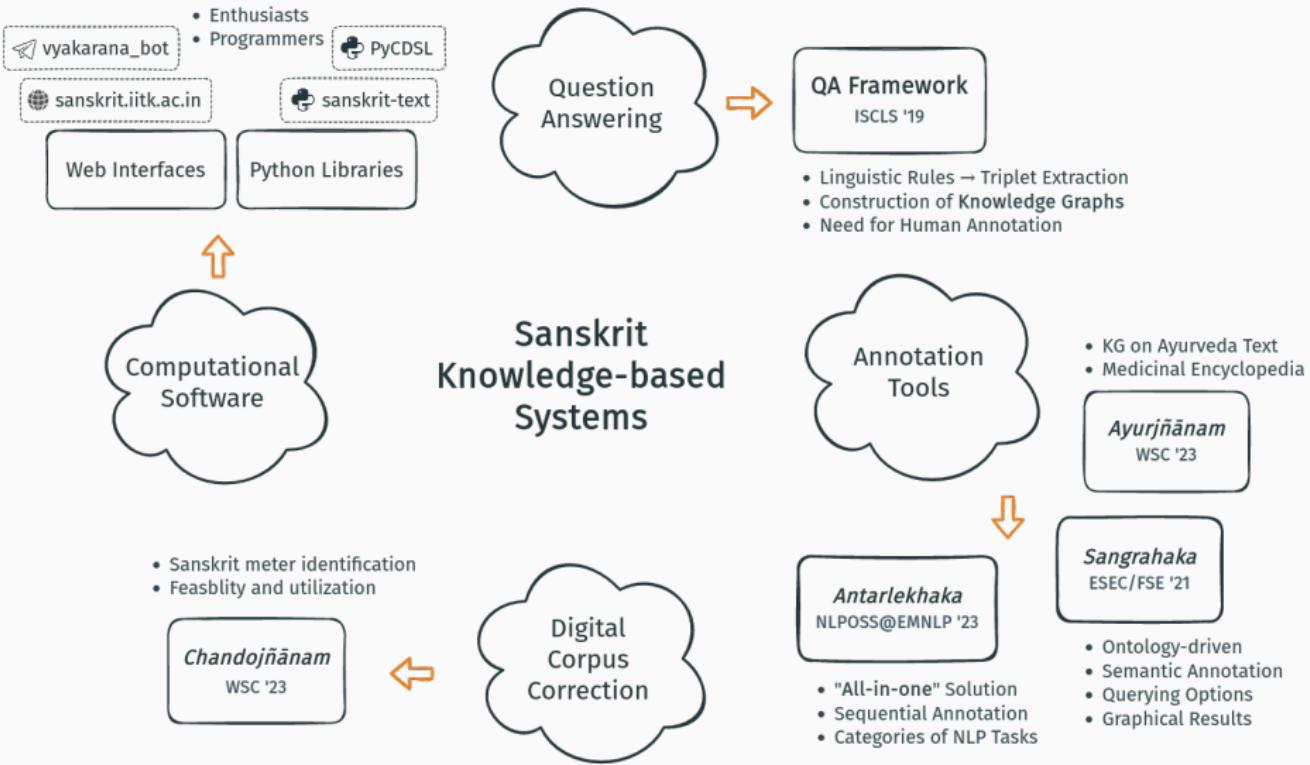
Entities detected by spaCy v3.4.4

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)



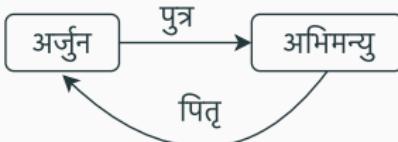
Problem

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 (mātṛ, mother), पुत्र (putra, son), etc.
- Multiple synonyms to the relationship words
 - पुत्री (putrī, daughter): दुहितृ (duhitṛ), तनया (tanayā), ...
- Implied Relationships

- Inverses



- Compositions



- Recursions



Framework

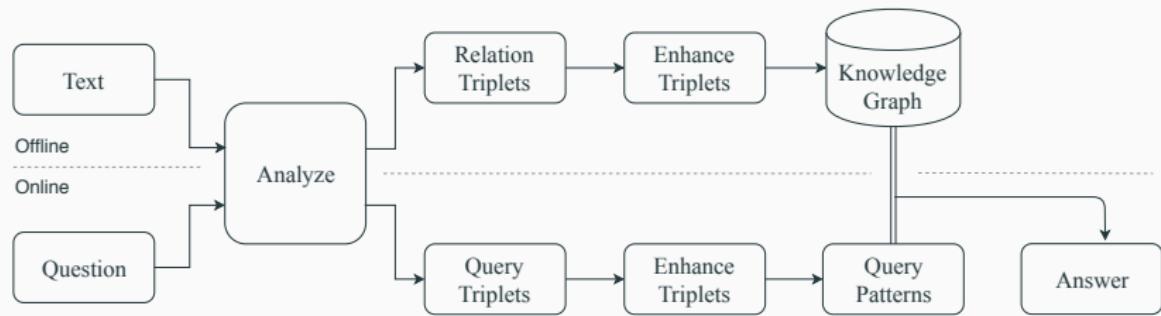
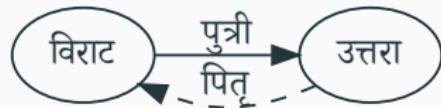


Figure 1: Overall framework of the Sanskrit QA system

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

पूरोभार्या कौसल्या बभूव।

(MBh 1.63.8c)

शर्मिष्ठायाः सुतो द्रुह्युस्ततोऽनुः पूरुरेव च॥

(MBh 1.79.21b)

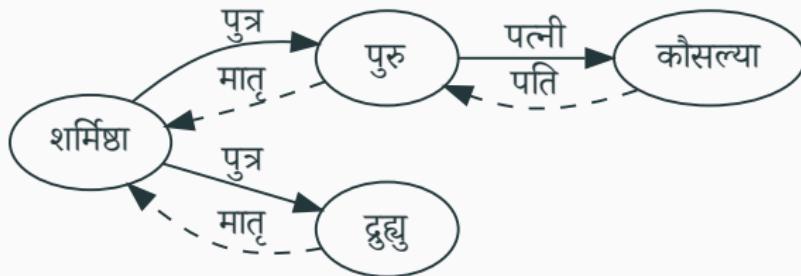
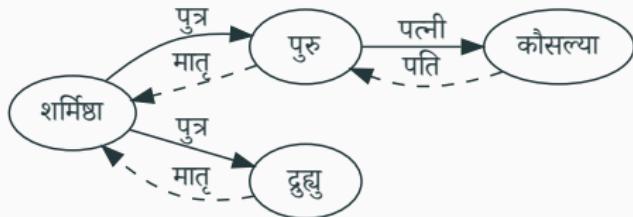


Figure 2: Knowledge graph enhanced with inverse relationships

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 का

(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	0.86
	QCond	60	55	31	0.56	0.46	0.50
	QAll	80	60	33	0.55	0.41	0.47

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 **saṃśāsa**
 - 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

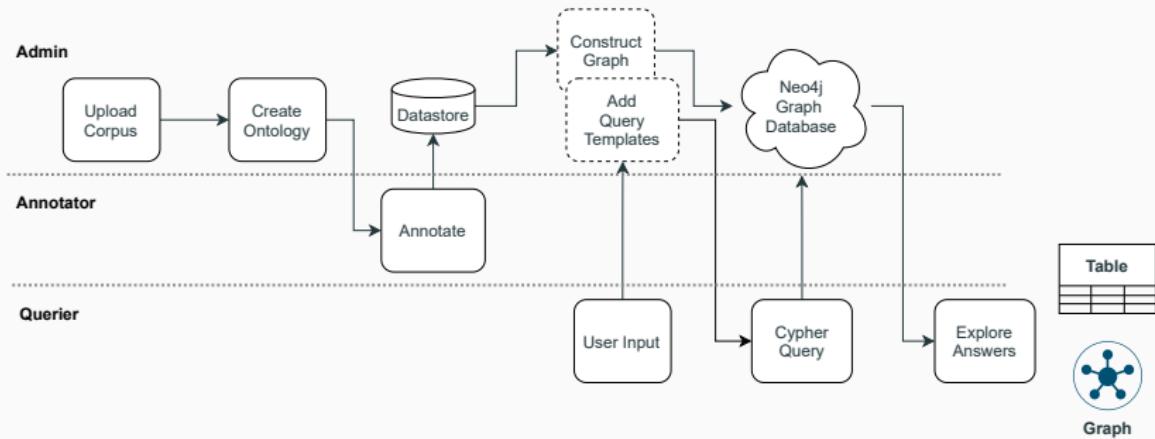


Figure 3: Architecture and Workflow of Sangrahaka

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



Preview

Bhavaprakasha Nighantu - यान्यर्थः

Line	Text				Split	?
256343	गोप्यः सुप्रतिपूर्विका शब्दसंकेतः				गोप्यः सुप्रतिपूर्विका शब्दसंकेतः	<input checked="" type="checkbox"/>
Word	गोप्यः	सुप्रति-	पूर्वि-	का		
Root	गोप्यः	सुप्रति-	पूर्वि-			
Gender	m.	n.		m.		
Case	1	2		2		
Number	sg.	sg.		pl.		
Noun?	true	true	false	false	true	false

Query

Show some details about #nil.

English

गोप्यः

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

Submit

Entity
Relation

Prepare

Line	256343
Entity	<input type="text"/>
Type	None

Prepare

Entities

✓ गोप्यः SUBSTANCE

✓ सुप्रति- SUBSTANCE

Confirm

Relations

✓ सुप्रति- :- [IS_SYNONYM_OF ()] → (गोप्यः)

Confirm

Query Result

x	a	p	r1	d2	x
गोप्यः	प्रति-	पूर्वि-	शब्दसंकेतः	गोप्यः	सुप्रतिपूर्विका शब्दसंकेतः
गोप्यः	प्रति-	पूर्वि-	शब्दसंकेतः	गोप्यः	गोप्यः
गोप्यः	प्रति-	पूर्वि-	शब्दसंकेतः	गोप्यः	गोप्यः

Showing 1 to 3 of 3 rows

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

Hrishikesh Rajesh Terdalkar

Sanskrit Knowledge-based Systems

20 / 48

Knowledge Graph on Bhāvaprakāśanighaṇṭu

Āyurjñānam – <https://sanskrit.iitk.ac.in/ayurveda/>

Corpus

- 2087 verses, 23 chapters¹
- Medicinal glossary

Ontology²

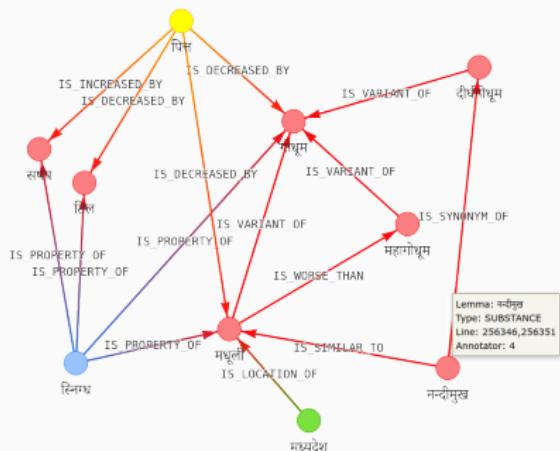
- 350+ entity types
- 430+ relationship types

Knowledge Graph

- 2400+ entities
- 5000+ relationships

Annotation Paradigm

- Capture semantic information
- Capture **unnamed entities**



¹We have annotated four chapters so far.

²We regularly update the ontology as and when required.

Semantic Information

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

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

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



Unnamed Entities

mudgo bahuvidhaḥ śyāmo haritaḥ pītakaḥ stathā.
śveto raktaśca teṣāntu pūrvah 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. rājikā \leftrightarrow kṣava \leftrightarrow kṣutābhijanaka \leftrightarrow kṛṣṇīkā \leftrightarrow kṛṣṇasarṣapa \leftrightarrow rājī \leftrightarrow kṣujjanikā \leftrightarrow āsurī \leftrightarrow tīkṣṇagandhā \leftrightarrow cīnāka
- Annotation: uṣṇa ⊢ is Property of → rājikā
- Query: Find all properties of cīnāka.
- **Problem:**
 - Relations might be connected to each other only in a chain
 - Potentially 10 edge traversal required!



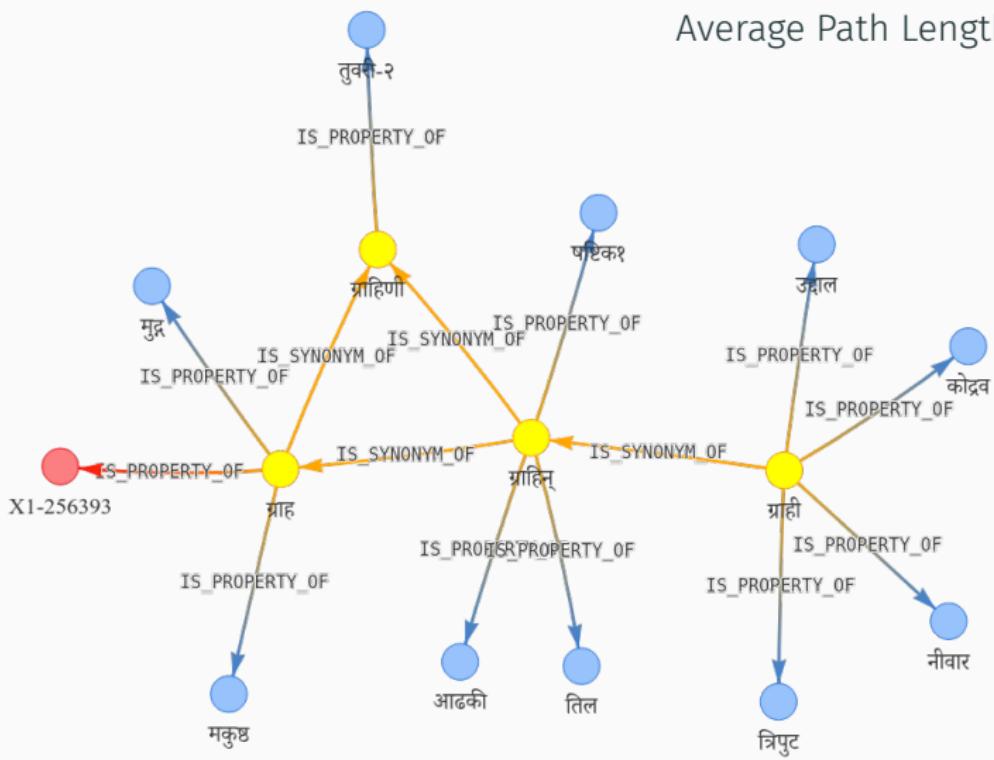
Synonyms – Solution

- 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.



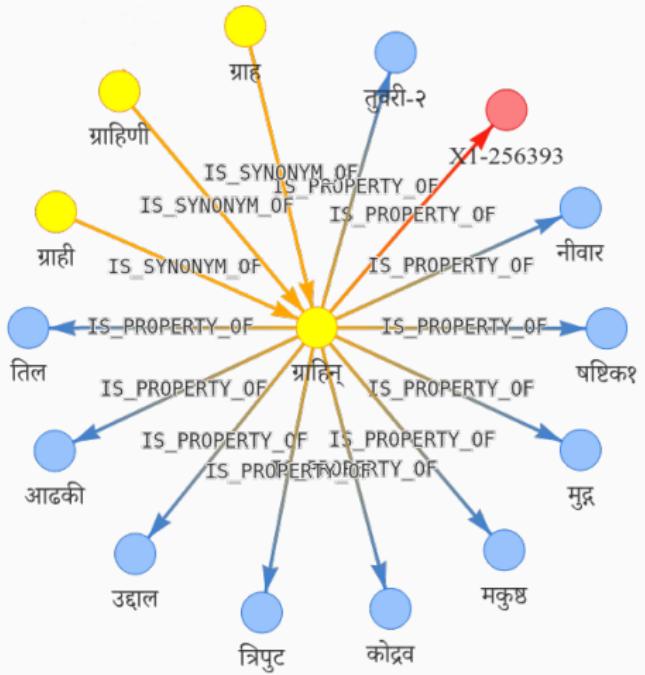
Example – Before Optimization

Average Path Length: 1.63



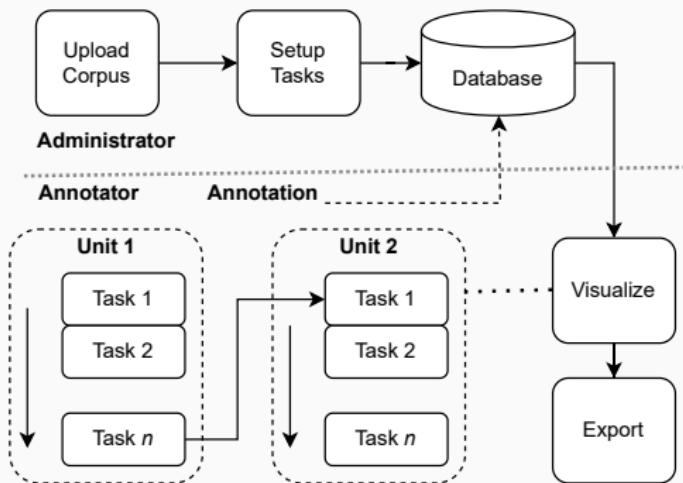
Example – After Optimization

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 rājyaśriyam gacchet] [striyā vākyavaśam gataḥ // 2
...]
[nāsyāparādhām paśyāmi] [nāpi doṣām tathāvidham] /
[yena nirvāsyate rāṣṭrādvanavāsāya rāghavaḥ] // 4



[ārye etad mama api na rocate]
[yad rāghavo rājyaśriyam 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.



Antarlekhaka – Scope

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	doccoano	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

Home Corpus Export Settings Admin Logout (Admin)

वाल्मीकिरामायणम्- Ayodhya 18

Search

Boundary Arvaya Lemma NER Action Coreference SentClass Discourse

Verse	Text	☰
○ 2488	तथा तु विलपन्ती तो कौसल्या राममातरम् उवाच लक्षणो दीनस्त तत् कालसदृशं वदः	■
○ 2489	न रोचते ममाय एतद् आर्यं यद् राघवो बनम् त्यक्त्वा राज्ञश्चियं गच्छेत् स्थिया वाक्यवर्णं गतः	■

Word	न	रोचते	ममाय	मम
Lemma	न	रुच्	-	मद्
UPOS	PART	VERB	-	PROI

Sentence Boundary

2488	तथा तु विलपन्ती तो कौसल्या राममातरम् उवाच लक्षणो दीनस् तत् कालसदृशं वदः ##
2489	न रोचते ममाय एतद् आर्यं ## यद् राघवो बनम् त्यक्त्वा राज्ञश्चियं गच्छेत् ## स्थिया वाक्यवर्णं गतः
2490	विपरीतश्च च वृद्धश्च च विषयेण च प्रधारिः गृष्म किम् द्वय न दूषण् चोहमानः सामग्र्यः ##
2489	

Submit

Evaluation

Subjective Evaluation

- Survey among annotators
 - Sangrahaka: 4.5/5
 - Antarlekhaka: 4.1/5



Figure 6: Wordcloud of testimonials

Objective Evaluation

- 29 Criteria
 - Technical
 - Functional
 - Data-related

Performance

- *Antarlekhaka*: 0.79
 - *Sangrahaka*: 0.74
 - INCePTION: 0.74
 - FLAT: 0.71
 - BRAT: 0.64



Tools and technologies

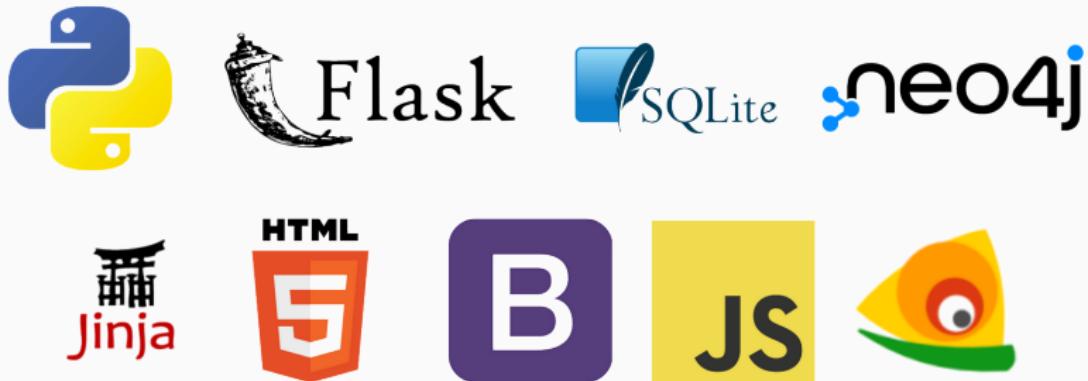


Figure 7: Tools and technologies that go into *Sangrahaka* and *Antarlekhaka*

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 signature³ 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/>

³ 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	✓ ⁴	✓ ⁵	✓	✓	✓
	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

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

⁵<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., वर्गै (vargai) as वर्गै (vagai)
 - Characters may be misidentified, e.g., ऋ (ṛ) as क्र (kra)
 - Characters may get split, e.g., ख (kha) as रव (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 धृतडारधनुर्धं बद्धपद्मासनस्थं
पीते वासो वसानं नवकमलदलस्तिष्ठेत्रं प्रसन्नम् ।
वामाङ्कासद्गं सीतामुखकमरमिलोचनं नीरदाभं
नानालङ्कारदीर्घं दध्मुरुजटामण्डनं रामचन्द्रम् ॥

Verse Mode Line Mode Identify

Results

Akṣarāṇī	ध्या ये दा जा नु बा हुं धृ त डा र ध नु षं ब द्ध प द्मा 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 मात्रा:
Jāti	प्रकृतिः
Chanda	सग्धरा (1 edit) + Fuzzy

Simulate digitization pipeline

- Generate PDF from Wikisource text
- Run two OCR systems: Google, Tesseract
- Obtain the OCR-ed versions of the text
- Three versions⁶ of **Meghadūta**⁷ 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

⁶Single text from different sources can differ in several places

⁷Also 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 Meters Identified	[Nei22] [Raj20] Chandojñānam	20 19 20	79 79 79	2 2 2	30 30 31	66 75 77	11 12 13	13 15 16	14 24 29	0 1 1	2 2 3	9 9 9	12 12 12	24 26 26	36 58 71	318 (80.3%) 364 (91.9%) 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 पर्वते ते (Pāda 3, Śloka 1.23)
- Incorrect word पर्वते (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', 'र(वे)[L]', 'ते'], ['ते']]]]
- Correctly points to the location where a change is required

Error #2

- Line: साभिज्ञानप्रहितकुशलैस्ततद्वचोभिर्ममापि (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
 - Bhūtasaṅkhyā: 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

- रामायणम्
- महाभारतम्
- भागवतप्रकाशनीयपदः
- श्रीमद्भगवतम्

Preferred number of words

Small

Encode number

14111265

Submit

Kaṭapayādi Encodings

- मां तारयत्यार्यः स्वल्पं
- मचकुकस्य यं वापि
- शचिष्ण्या कार्यं त्वया

Decode text

मां तारयत्यार्यः स्वल्पं

Submit

Kaṭapayādi Number

14111265

Split	म	ा	ं	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

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

Features

- Stem finder (**Prātipadikam**)
 - Declension generator (**Subantāḥ**)
 - Root finder (**Dhātuḥ**)
 - Conjugation generator (**Tiñantāḥ**)
 - Word segmentation (**Sandhisamāsau**)



```
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()

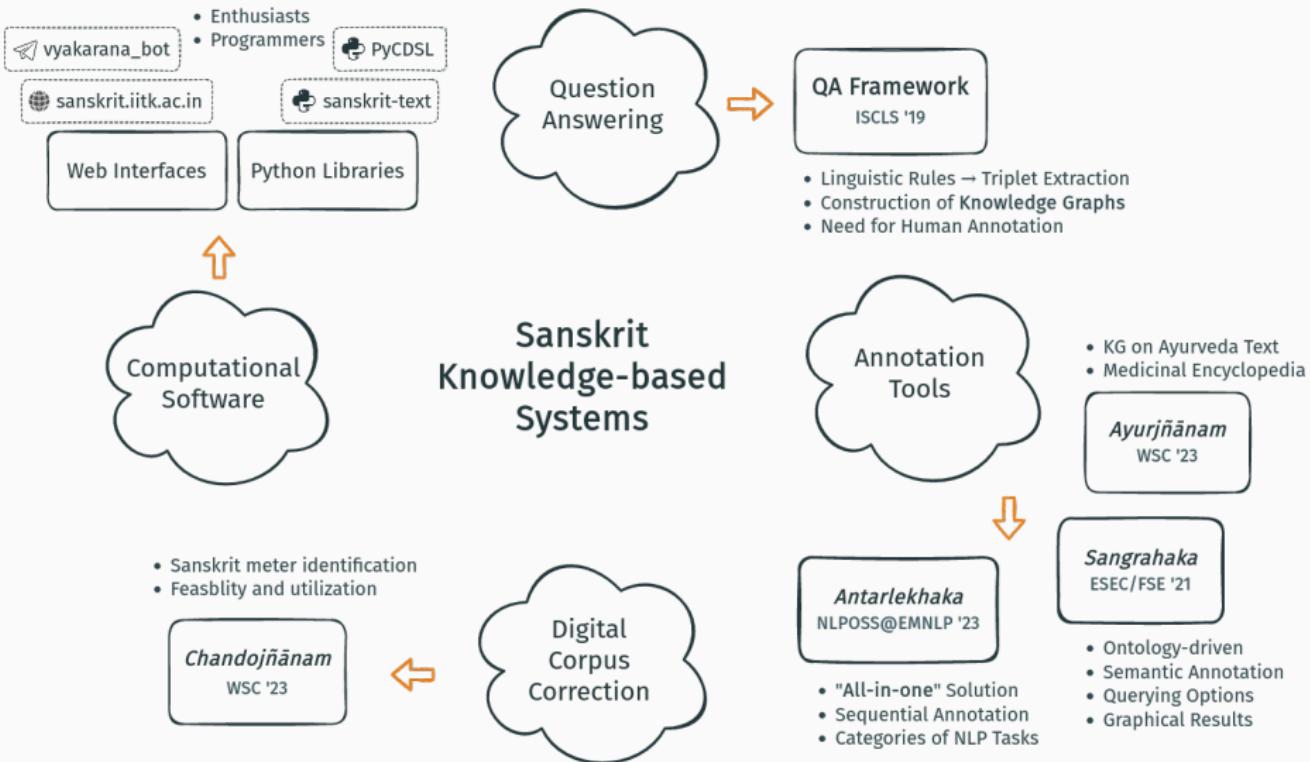
# dictionary accessible using `[]` operator
results = CDSL["MW"].search("राम")

# 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



Future Work

- 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

-  Hrishikesh 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.
-  Hrishikesh Terdalkar and Arnab Bhattacharya.
Sangrahaka: 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.
-  Hrishikesh 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, Hrishikesh Terdalkar, Tushar Sandhan, Suvendu 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.
-  Hrishikesh Terdalkar and Arnab Bhattacharya.
Chandojananam: 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.
-  Hrishikesh 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: कर्णजुनयोः कश्श्रेष्ठः (Who was greater – Karṇa or Arjuna?)
- Splitting of *samāsa* and *sandhi*
 - *Sanskrit Sandhi and Compound Splitter*⁸
 - Output: कर्ण-अर्जुनयोः कः-श्रेष्ठः
- Morphological analysis of the word
 - *The Sanskrit Heritage Platform*⁹
 - 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.}

⁸Oliver Hellwig, Sebastian Nehrdich: *Sanskrit Word Segmentation Using Character-level RNN and CNNs*. EMNLP 2018.

⁹The 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 (*śaṣṭhī 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 (**samyogah**)
 - 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, Chandovicitih, 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