



INDIAN KNOWLEDGE SYSTEMS DIVISION

Ministry of Education, Govt of India.

AICTE Headquarters, Nelson Mandela Marg, Vasant Kunj, New Delhi-110070.

Website: www.iksindia.org Tel: 011-29581523/1004.

PROJECT COMPLETION REPORT

Important Points:

- i. The Project Completion Report (PCR) should cover the complete work done during the entire project.
- ii. Timely submission of report is essential to facilitate the release of funds.
- iii. Report should be in the format given below.
- iv. Write "NIL" against items where there is nothing significant to report or if these items are not relevant.
- v. One soft copy in .pdf format of the report giving information on items listed in the format given below are required. There is no requirement to send a hardcopy.

PROJECT COMPLETION REPORT

PROJECT TITLE

DEVELOPMENT OF CONVERSATIONAL SANSKRIT TRANSLATOR USING
NEURAL MACHINE TRANSLATION

SUBMITTED BY

Prof. ANIL KUMAR GOURISHETTY

Indian Institute of Technology, Roorkee

20 December 2024

FUNDED BY

Indian Knowledge Systems Division, Ministry of Education,
Govt of India

IKS Division File Reference No: AICTE/IKS/REFP1/2021-22/05 dated April 28,
2022

Project Duration: From Dec 2021 to Dec 2024

PROJECT COMPLETION REPORT

One soft copy of the Project Completion Report should be submitted on email. The Report should consist of Part A - Summary Report and Part B - Comprehensive Report. The report should be on the Standard Format provided.

1. Project Title: DEVELOPMENT OF CONVERSATIONAL SANSKRIT TRANSLATOR USING NEURAL MACHINE TRANSLATION

2. Duration of Project: From Dec 2021 To Dec 2024

3. Focus Area: Machine Translation (Sanskrit-English and Sanskrit-Hindi)

4. Principal Investigator

Name: Prof. ANIL KUMAR GOURISHETTY

Organisation/Institution: IIT Roorkee

5. Co-Principal Investigators

Name: Prof. Ganesh Ramakrishnan

Organisation/Institution: IIT Bombay

Name: Dr. Jitin Singla

Organisation/Institution: IIT Roorkee

6. Collaborating Organisations

Organisation	Team Members	Activities/Tasks assigned
IIT Roorkee	S Keerthana (Intern), Mridul Sharma, Arun Mampazhy (Project Manager)	Translation, reviews, model training and testing, Project management
IIT Bombay	Ayush Maheshwari, Anuja Dumada, Rahul Kharat, Pravin Ingle, Karthika N J.	Model training, Testing, SSKLS tool development, transition
KSU	Chaitanya L, Aatmaram Bhat.	Translation, reviews in SSKLS Tool

Part A - Summary Report

1. Project Objectives

Objective as per the approved Project	Fully Achieved/Partially Achieved (indicate shortfall)	Reason for Partial Achievement
Generate a corpus with a large number of Sanskrit-to-Hindi and Sanskrit-to-English sentences	Yes, details in deliverables section	
Develop a Sanskrit-to-Hindi translator using Neural Machine Translation	Yes, details in deliverables section	
Develop a Sanskrit-to-English translator using Neural Machine Translation	Yes, details in deliverables section	

2. Deliverables

Deliverables as per the approved Project	Fully/Partially/Not Achieved	Reason for Partial/ Non Achievement
Sanskrit-English and Sanskrit- Hindi Parallel datasets for Machine Translation, including Benchmark data	About 53,000 parallel sentences collected and processed for Sanskrit-English 91,855 parallel sentences collected and processed for Sanskrit-Hindi. (Approx. NIOS 49000, Spoken tutorials 25000, Chandamama 9000, Mann-ki-baat 6000 and MKGandhi - 2000)	Nil
Translation models trained to translate between San-Hin and San-Eng pairs	MT models trained with our dataset - using the base models IndicTrans2, NLLB, IndicBART, ByT5	-

3. Specific Benefits/Outcome

(Please give details)

- Patent, if any: Nil
- Product/Process developed/ technology transferred

Translation models trained to translate between San-Hin and San-Eng pairs.

- List of Publications arising from the Project

(Indicate Impact Factors of the Journals; attach Copies of the Papers)

Samayik: A Benchmark and Dataset for English-Sanskrit Translation

Based on the final phase work, future publication(s) may be there

- Linkages developed

4. Manpower trained

- No. of Students based on the project: 10 students from IIT Roorkee and KSU
- No. of Master's Thesis/Reports based on the project: 1 (Keerthana)
- Position and Number of Other Technical Staff Trained: 1 technical staff - Internship, continued as Project Associate

5. Methodology (Minimum 250 Words):

- Salient Research Achievements

Created parallel datasets for the language pairs Sanskrit-Hindi and Sanskrit-English with a total of 53,000 and 91,855 sentences respectively.

Developed a Sanskrit translation system, for an automatic translation of conversational Sanskrit sentences to Hindi and English. Sanskrit dramatically differs from English as well as from Hindi which are two prominent languages in use in India today. Sanskrit is a grammatically rich language and possesses free word order whereas languages such as English follows a Subject-Verb-Object order and Hindi follows a Subject-Object-Verb order. Due to this, Sanskrit translation is a very challenging task and thus requires a very large corpus of existing translations to train an efficient machine learning based translator. In addition to this, translation of Sanskrit to English is further challenging due to completely different grammatical structures. We developed 4 translation systems English to Sanskrit, Sanskrit to English, Hindi to Sanskrit, and Sanskrit to Hindi. These systems are trained on the parallel data prepared by us for the respective language pairs.

- New Observations (minimum 200 words):

Machine translation quality of both the language pairs Sanskrit - English and Sanskrit-Hindi, in both directions, performed using various MT models (NLLB, IndicTrans, ByT5) shows consistent improvement with our datasets used as training data for the respective languages.

- Innovations (Give bulleted list):

- Creation of data sentences in simple Sanskrit
- Development of two MT models

6. Indication of scope for future work (Maximum 300 Words):

Good quality data is an integral part of any ML-based Natural Language Processing tasks. With the new, good quality parallel datasets created, for the 2 sets of language pairs, can contribute to the community for various multilingual or cross lingual NLP tasks concerning these languages. Sanskrit, being a grammatically well-structured language, and the majority of Indian languages having a large overlap in the vocabulary, we expect the datasets as well as the models trained on them, to further aid in the tasks involving other Indian languages.

We also consider the 2 sets of language pairs considered are very important for the further research (i) Sanskrit- English : Text contents in English forms the single largest digitally available data. Parallel sentences with English as one of the languages can thus aid in various NLP tasks involving multilinguality such as cross-lingual transfer learning. (ii) Sanskrit – Hindi: Among all Indian languages, the digital data availability is the highest in Hindi. Given the commonality in both grammatical structure and vocabulary overlap in Indian languages, prior works have shown that Hindi can act as a pivot language in various cross-lingual NLP tasks.

Sanskrit, being more structured, and also sharing a larger subset of vocabulary (either directly or derived) with most major Indian languages, can additionally help in such tasks, forming a well-structured pivot language.

7. Procurement/Usage of Equipment:

Sr.	Name of Equipment	Make/ Model	Cost (Rs.)	Date of Purchase & Installation	Remarks regarding maintenance/ breakdown
1	7000 Desktop	Dell Optiplex	1,42,120	15/11/2022 & 05/12/2022	
2	Projector	EPSON /EP-E01	42,240	03/11/2022 & 03/11/2022	
3	Printer supply	HSN-9010	5,900	03/11/2022 & 03/11/2022	

Part B – Comprehensive Report

(The Comprehensive Report should be precise in detail and self-contained)

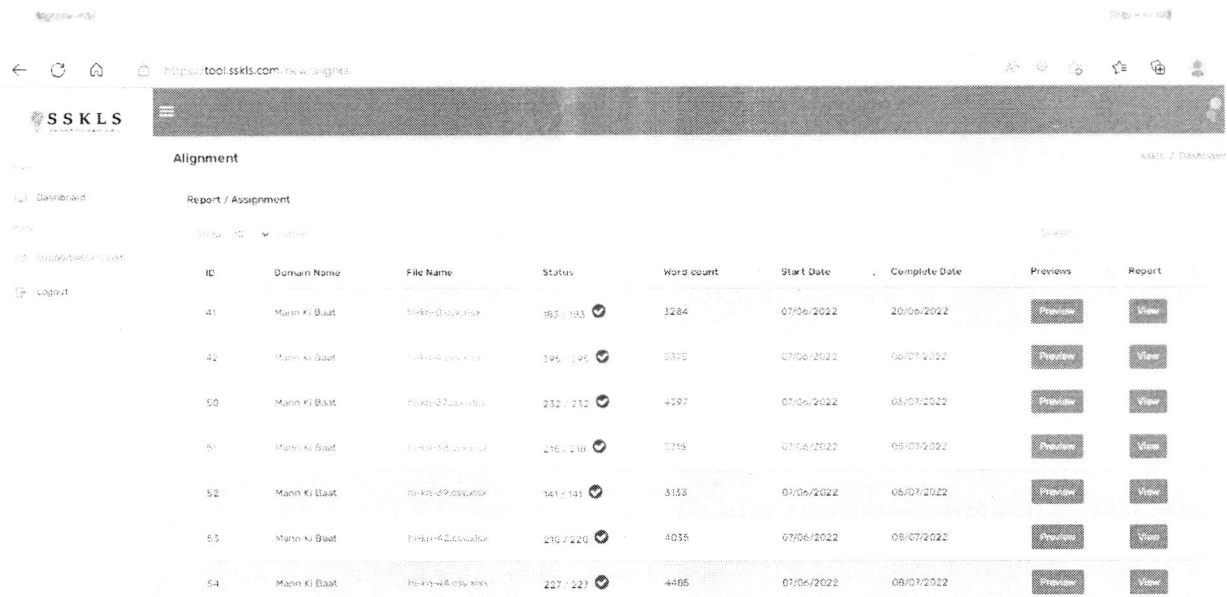
1. Project Title: Development of Conversational Sanskrit Translator Using Neural Machine Translation

2. Literature Review

- Identify gaps in knowledge that your project aims to address
- Explain how your research builds upon or contributes to existing knowledge

3. Methodology adopted for Development of the Product/Process in the Research Project *(State briefly in 800 words)*

The development of the SSKLS tool for corpus creation followed a systematic methodology to address the challenges of annotating text data, particularly for low-resource Indian languages. The process began with problem identification and requirement analysis, followed by a review of existing tools to identify gaps. The tool was designed with a modular architecture, integrating a robust preprocessing pipeline for multilingual data, a keyword-based semi-supervised learning engine, and an interactive annotation interface. Rigorous testing using benchmark datasets ensured high annotation quality and iterative improvements were made based on user feedback. The resulting tool significantly reduced manual effort, enhancing the efficiency and scalability of corpus creation for diverse linguistic applications.



The screenshot displays the SSKLS web application interface. The top navigation bar includes the SSKLS logo and a menu icon. The main content area is titled "Alignment" and shows a table of data. The table has columns for ID, Domain Name, File Name, Status, Word count, Start Date, Complete Date, Previews, and Report. The data rows show various files with their respective word counts and dates. The interface also includes a sidebar with navigation links like Dashboard, Reports, and Logout.

ID	Domain Name	File Name	Status	Word count	Start Date	Complete Date	Previews	Report
41	Marathi Baat	Baat-01.docx	103 / 103 ✓	3284	07/06/2022	20/06/2022	Preview	View
42	Marathi Baat	Baat-02.docx	195 / 195 ✓	5375	07/06/2022	06/07/2022	Preview	View
50	Marathi Baat	Baat-03.docx	232 / 232 ✓	4297	07/06/2022	05/07/2022	Preview	View
51	Marathi Baat	Baat-04.docx	216 / 216 ✓	5715	07/06/2022	05/07/2022	Preview	View
52	Marathi Baat	Baat-05.docx	141 / 141 ✓	3133	07/06/2022	05/07/2022	Preview	View
53	Marathi Baat	Baat-06.docx	210 / 220 ✓	4035	07/06/2022	08/07/2022	Preview	View
54	Marathi Baat	Baat-07.docx	227 / 227 ✓	4465	07/06/2022	08/07/2022	Preview	View

4. Description of any Product/Process developed during the project

5. Outputs & Outcomes:

a. Technology Development

- Technology developed at Lab scale/pilot scale/commercial scale: Translation models trained on the data created by the team
- Technology demonstration in field setup - Yes/No No
- Technology Transfer to industry: If Yes (Name of the Industry) Nil
- Buyers/End users of product: Nil
- Affordability: The cost estimate of the developed prototype

b. Knowledge Creation

- Patent filings: Nil
- Participation in conference/ training workshops attended:
LREC-COLING 2024 (Virtual mode - presentation of Samayik paper)

c. Knowledge Dissemination

- Describe the strategies used to disseminate the research findings and knowledge. • Examples of dissemination methods:

i. Conference presentations

LREC-COLING 2024 (Virtual mode - presentation of Samayik paper)

ii. Workshops and training sessions None

iii. Online resources (websites, open access repositories)

Code and dataset are available at <https://github.com/ayushbits/samayik>

iv. List of Courses developed None

v. Social media/other media outreach -

d. Publications

Type of Publications	Number of Publications	List of Publications	Authors & Co Authors	Theme of the publication	References/links of the Publications (link not required for books)
Research Papers Additional publication on Sanskrit-Hindi work is in discussion	1	Sāmayik: A Benchmark and Dataset for English-Sanskrit Translation	Ayush Maheshwari, Ashim Gupta, Amrith Krishna, Atul Kumar Singh, Ganesh Ramakrishnan, G. Anil Kumar, Jitin Singla	Release of parallel dataset for English-Sanskrit sentences written in contemporary prose.	https://aclanthology.org/2024.lrec-main.1245.pdf

Books	In discussion				
Articles	NIL				
Poster presentations	NIL				
Journals published through IKS center / Project	NIL				
Magazines	NIL				

e. Describe how the project How the project resulted in integration of IKS and Contemporary knowledge

The project aims to connect Sanskrit which is at the core of IKS with the most spoken Indian language as well as English using modern technology

f. Knowledge Repository

- Describe the structure and organization of the knowledge repository

We've created datasets of parallel sentences for Sanskrit - English and Sanskrit - Hindi language pairs. About 53000 parallel sentences collected and processed for Sanskrit-English, 91,855 parallel sentences collected and processed for Sanskrit-Hindi.

- Explain how users/public can access and utilize the knowledge repository.

The dataset for Sanskrit-English is publicly available and accessible, along with code to train and test various widely used Machine Translation models at the following link
<https://github.com/ayushbits/saamayik>

g. Capacity building

- Officers/Faculty
 - Temporary manpower recruited/trained
- 1 technical staff worked as part of the project full-time as an Intern and continued with the project as a Project Associate

h. Results & Analysis with figures

- Present your findings in a clear and organized manner
- Include figures, Data analysis (Tables, Charts, Graphs), and Illustrations to support your findings.

Overall data statistics and translation results

1. Sanskrit - English

Dataset	NIOS	Spoken Tutorials	GitaSopanam	Bible	Mann Ki Baat	Total
#sentences	11356	23835	5885	7838	4047	52941
#words	105178	237449	26135	102508	47843	518842
#unique words	30966	38373	6513	38359	20484	122349
% of unique words	29.4	16.2	24.9	37.4	42.8	23.6
Mean word length	9.3	10	4.5	13.1	11.8	9.8

Table 1: Number of sentences, words, unique words and average word length for different corpus in Sāmāyik.

	En-Sa		Sa-En	
Model	BLEU	ChrF	BLEU	ChrF
ByT5	28.7	44.4	31.1	55.7
mBART	27.20	46.61	11.6	27.0
IndicBART	25.45	43.47	29.79	50.14
Indictrans	11.3	46.6	37	58.2

Table 2: Results for different models on the in-domain test set for En-Sa and Sa-En direction.

	En-Sa		Sa-En	
Model	BLEU	ChrF	BLEU	ChrF
ByT5	7	21.4	5.4	29
mBART	7.11	22.6	-	-
IndicBART	6.9	22.4	5.3	27.7
Indictrans	0.6	26.7	13.1	37.5
Google Trans	1.9	35	13.9	44.7
NLLB	1.2	27.6	11.5	36.1
Indictrans(Vanilla)	1.2	34	14.5	42.7

Table 3: Results for out-of-domain test set, namely, Mann Ki Baat (MKB) for En-Sa and Sa-En directions. We omit mBART due to poor performance on in-domain test split (refer Table 2). Performance reported on the Google Translate, NLLB and Indictrans (Vanilla) (below double horizontal line) refers to the evaluation with pre-trained models.

Model	IndicBART		mBART	
Dataset	BLEU	ChrF	BLEU	ChrF
Itihasa	4.6	16	4.3	16.7
Sāmāyik	6.3	23.2	7.3	22.3
Itihasa + Sāmāyik	6.9	22.4	6.8	21.6

Table 4: Comparison between existing dataset Itihasa, Sāmāyik and Itihasa + Sāmāyik on MKB out-of-domain testset for English-Sanskrit translation direction. The score difference between Itihasa (1st row) and Sāmāyik(2nd row) are statistically significant at $p < 0.05$.

i. Achievements/ highlights of the project

- Datasets created and(or) curated

About 53,000 parallel sentences collected and processed for Sanskrit-English

91,855 parallel sentences collected and processed for Sanskrit-Hindi.

- Translation models trained on the above datasets for both the pairs of Hindi-Sanskrit and Sanskrit - English, in both directions.

j. References

- Popovic, Maja. "chrF: character n-gram F-score for automatic MT evaluation." WMT@EMNLP (2015).
- [Bleu: a Method for Automatic Evaluation of Machine Translation](https://aclanthology.org/P02-1040) (Papineni et al., ACL 2002)
- <https://huggingface.co/spaces/evaluate-metric/bleu>
- <https://huggingface.co/spaces/evaluate-metric/chrf>
- https://huggingface.co/docs/transformers/model_doc/nllb
- <https://huggingface.co/ai4bharat/IndicBART>

k. Appendices

- Include any additional information that supports the research, may not fit within the main body of the report.

l. Project Outcomes Contribution to (Bullet points. No more than 250 words in each category)

- **Dr̥ṣṭi:** Sanskrit in itself is the core "Indian" element here and this project contributes to connecting and enhancing data
- **Jñāna:** Data used is a mix of what can be called ancient and modern, but in Sanskrit, contributing to the Jnana-Parampara
- **Laukika-prayojana:** Connecting more people to India's knowledge treasure through a language that is at its core by using, developing and enhancing tools and methods that are useful to the present community

Name of PI
(Anil Kumar Gourishetty)

Name of Co-PI-1
(Ganesh Ramakrishnan)

Name of Co-PI-2
(Jitin Singla)

Signature of PI

Signature of Co-PI-1

Signature of Co-PI-2

Head of the institution
Dean
Sponsored Research & Industrial Consultancy
Indian Institute of Technology Roorkee
Roorkee-247 667 (INDIA)

