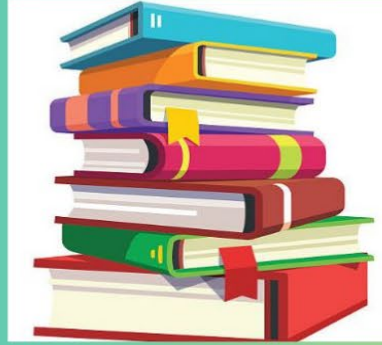


ప్రామాణిక తెలుగు సహజ పాఠ్య దత్తాంశం

సంపుటం- ౨



**A GOLD STANDARD TELUGU RAW
TEXT CORPUS VOL. II**

A GOLD STANDARD TELUGU RAW TEXT CORPUS VOL. II



34

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

**Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysuru, India-570006**

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Manasagangotri, Mysuru, Karnataka, India, 570006
www.ciil.org

A Gold Standard Telugu Raw Text Corpus Vol. II

Dr. Modugu Kasimbabu, Rajesha N., Manasa G., Dr. Narayan Kumar Choudhary,
Prof. Shailendra Mohan

e-ISBN: 978-93-48633-12-5

CIIL Publication No.: 1511

First published: AD 2025 March :: Phalguna 1946 Śaka

© Central Institute of Indian Languages, Mysuru 2025

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit

Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit

B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit

M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Umesh Chamling Rai

TABLE OF CONTENTS

Table of Contents	iv
Figures	iv
Tables	iv
1 A Gold Standard Telugu Raw Text Corpus Vol. II	1

FIGURES

Figure 1: Representation of the Domains in Telugu Raw Text Corpus Vol. II	2
Figure 2: Representation of the subdomains in Telugu Raw Text Corpus Vol. II	3

TABLES

Table 1: Representation of the Domains in Telugu Raw Text Corpus Vol. II	1
Table 2: Representation of the Sub-domains in Telugu Raw Text Corpus Vol. II	2

1 A GOLD STANDARD TELUGU RAW TEXT CORPUS VOL. II

1.1 INTRODUCTION

Linguistic Data Consortium for Indian Languages (LDC-IL) released “A Gold Standard Telugu Raw Text Corpus” in 2018 [1], which was developed from available contemporary texts. The corpus creation followed the pre-defined criteria which are mentioned at [2]. This corpus contains 30,10,993 words, and character count is 2,49,14,821 sourced from sourced from books, newspapers, and magazines, that can be refer at [3]. Telugu text corpus is collected from various libraries in Andhra Pradesh, mostly from Hyderabad, Vishakahpattanam, Kuppam, Guntoor, Thirupathi and Ananthpur. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, and economics have very less amount of books. Literary texts are easily available in Telugu but getting scientific text is very difficult. Some categories like epigraphy, finance, Commerce, oceanology text are rare in these libraries [4]. It provides comprehensive representations of major domains such as Aesthetics and Social Sciences, among others. The Telugu Raw Text Corpus, in its original form, appears to be relatively small, limiting its effectiveness for comprehensive linguistic analysis and Natural Language Processing tasks. To address this limitation LDC-IL has made significant enhancements to the corpus by incorporating additional domains of text. This expansion not only increases the overall size of the corpus but also diversifies its content, making it more representative of the variety of language use across different contexts.

1.2 TELUGU RAW TEXT CORPUS VOL. II

A Gold Standard Telugu Raw Text Corpus Vol.II is a comprehensive collection of Telugu language texts, comprising 30,13,530 words drawn from different titles. This corpus includes extracts from Govt. Text books, books, Govt. official documents, news, and magazines providing a diverse range of linguistic data. It can be used as a valuable resource for linguistic research, language processing applications, and the preservation of the Telugu language. This volume has the representation of six major domains covered that are Aesthetics, Commerce, Mass Media, Official Documents, Science and Technology and Social Sciences. Researchers and developers can utilize this resource to enhance their understanding and applications related to the Telugu language

The representation of the six major domains covered has been shown in the table below:

Domain	Word count
Aesthetics	2,55,289
Commerce	1,77,336
Mass Media	6,02,327
Official Document	4,06,454
Science and Technology	6,71,135
Social Sciences	9,00,989
Total	30,13,530

Table 1: Representation of the Domains in Telugu Raw Text Corpus Vol. II

The representation of the sub-domains covered has been shown in the table below:

Category	Sub Category	Word Count
Aesthetics	Folklore	14809
Aesthetics	Literary Texts	29425
Aesthetics	Literature-Criticism	41988
Aesthetics	Literature-Text Books (School)	169067
Commerce	Accountancy	86467
Commerce	Business	80027
Commerce	Finance	10842
Mass Media	General News	367992
Mass Media	Political	227835
Mass Media	Social	6500
Official Document	Administration	249311
Official Document	Legislature	53271
Official Document	Parliamentary/Assembly Debates	23131
Official Document	Police Documents	80741
Science and Technology	Biology	90521
Science and Technology	Educational Psychology	58588
Science and Technology	Environmental Science	37292
Science and Technology	Mathematics	42787
Science and Technology	Naturopathy	18082
Science and Technology	Physics	81381
Science and Technology	Text Book (Science)	342484
Social Sciences	Economics	51969
Social Sciences	Education	192386
Social Sciences	Geography	12532
Social Sciences	History	33396
Social Sciences	Physical Education	103060
Social Sciences	Political Science	65640
Social Sciences	Sociology	137074
Social Sciences	Text Book (Social Science)	304932

Table 2: Representation of the Sub-domains in Telugu Raw Text Corpus Vol. II

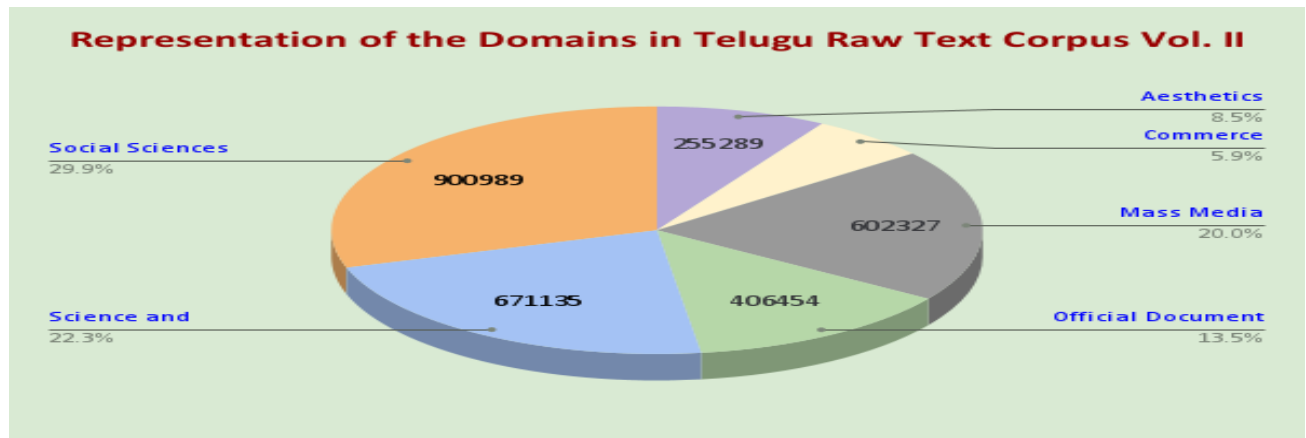


Figure 1: Representation of the Domains in Telugu Raw Text Corpus Vol. II

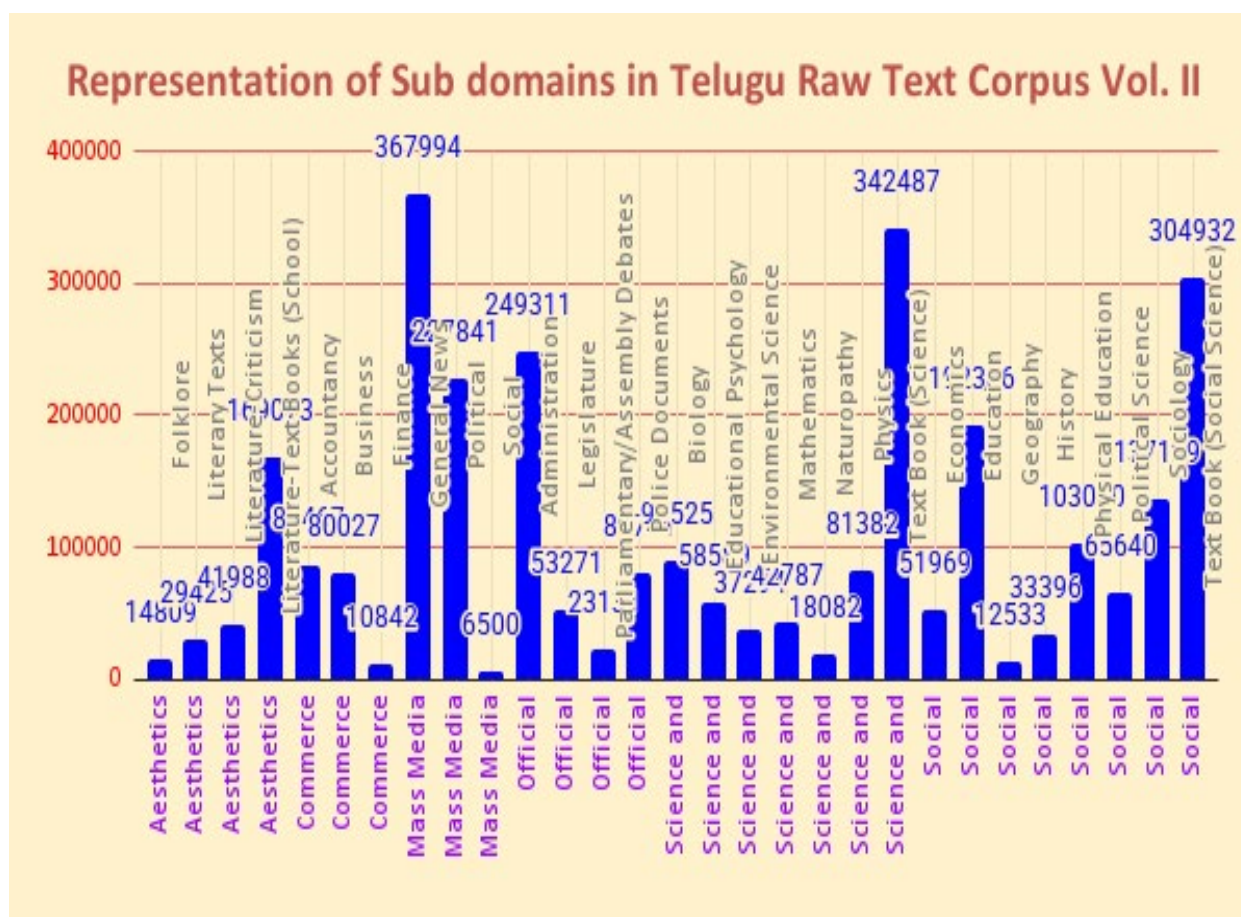


Figure 2: Representation of the subdomains in Telugu Raw Text Corpus Vol. II

1.3 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora in Proceedings of 23rd Conference of the Oriental COCOSA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSA), Yangon, Myanmar, 2020. pp. 28-32, doi: <https://doi.org/10.1109/O-COCOSA50338.2020.9295011>
2. Choudhary, N. 2021. LDC-IL: The Indian Repository of Resources for Language Technology. Language Resources & Evaluation. Springer, Vol. 55, Issue 1. doi: <https://doi.org/10.1007/s10579-020-09523-3>
3. Choudhary, Narayan, Rajesha N., Manasa G. & L. Ramamoorthy. 2019. "LDC-IL Raw Speech Corpora: An Overview" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 160-174.
4. Ramamoorthy, L., Narayan Choudhary, Thirupal C Reddy & Gangaraju H. 2019. A Gold Standard Telugu Raw Text Corpus Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Sajila S., Rajesha N., Manasa G., 2018. "Telugu Raw Text Documentation" A Gold Standard Telugu Raw Text Corpus in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-271-2.