

معیاری کأشُر خام متتک کارپس جلد ۲

A GOLD STANDARD

KASHMIRI

RAW TEXT CORPUS VOL.II

A GOLD STANDARD KASHMIRI RAW TEXT CORPUS VOL.II

A GOLD STANDARD KASHMIRI 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 Kashmiri Raw Text Corpus Vol. II

Dr. Zargar Adil Ahmad, Bi Bi Mariyam, Rajesha N., Manasa G., Dr. Narayan Kumar Choudhary,
Prof. Shailendra Mohan

e-ISBN: 978-93-48633-27-9

CIIL Publication No.: 1510

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: Zargar Adil Ahmad

TABLE OF CONTENTS

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

FIGURES

Figure 1: Representation of the Domains in Kashmiri Raw Text Corpus Vol. II	2
---	---

TABLES

Table 1: Representation of the Domains in Kashmiri Raw Text Corpus Vol. II.....	1
Table 2: Representation of Sub domains in Kashmiri Raw Text Corpus Vol. II	4

1 A GOLD STANDARD KASHMIRI RAW TEXT CORPUS VOL. II

Dr. Zargar Adil Ahmad

1.1 INTRODUCTION

In 2019, the LDC-IL released a Gold Standard Kashmiri Raw Text Corpus [1], which was developed from available contemporary texts. The corpus creation followed the pre-defined criteria which are mentioned at [2]. This corpus contains 466,054 words and a character count of 2,646,948, sourced from books, newspapers, and magazines that can be refer at [3]. It provides comprehensive representations of major domains such as aesthetics and social sciences, among others. The Kashmiri 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 KASHMIRI RAW TEXT CORPUS VOL. II

A Gold Standard Kashmiri Raw Text Corpus Vol. II is a comprehensive collection of Kashmiri language texts, comprising 10, 13,658 words and 57, 28,547 characters. This corpus includes extracts from books, newspapers, and magazines, providing a diverse range of linguistic data. It serves as a valuable resource for linguistic research, language processing applications, and the preservation of the Kashmiri language. This volume has the representation of six major domains covered as compared to previous volume which has only two major domains of Aesthetics and social sciences. Researchers and developers can utilize this resource to enhance their understanding and applications related to the Kashmiri language.

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

Domain	Word count	Percentage
Aesthetics	2,67,140	76.93%
Commerce	1,77,666	0.28%
Mass Media	6,18,096	10.50%
Official Document	4,06,791	0.55%
Science and Technology	7,27,632	3.00%
Social Sciences	9,36,637	8.72%
Total	10,13,658	100%

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

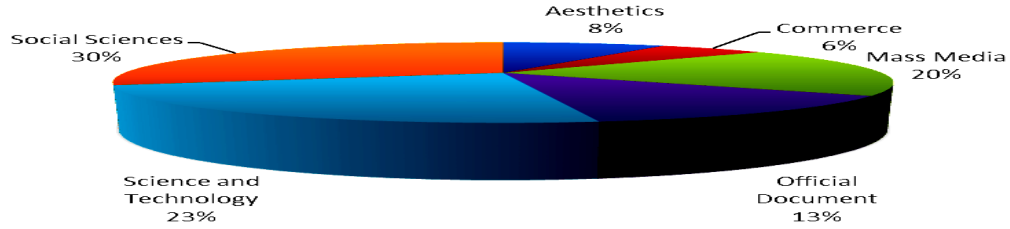


Figure 1: Representation of the Domains in Kashmiri Raw Text Corpus Vol. II

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

Category	Sub Category	Word Count	Percentage (within Sub domain)	Overall Percentage
Aesthetics	Biographies	28915	3.71%	2.85%
Aesthetics	Culture	12930	1.66%	1.28%
Aesthetics	Fine Arts-Dance	288	0.04%	0.03%
Aesthetics	Folk Tales	7329	0.94%	0.72%
Aesthetics	Literary Texts	324660	41.63%	32.03%
Aesthetics	Literature-Children's Literature	1267	0.16%	0.12%
Aesthetics	Literature-Criticism	223237	28.63%	22.02%
Aesthetics	Literature-Epics	840	0.11%	0.08%
Aesthetics	Literature-Letters	3960	0.51%	0.39%
Aesthetics	Literature-Novels	62589	8.03%	6.17%
Aesthetics	Literature-Plays	24905	3.19%	2.46%
Aesthetics	Literature-Poetry	1383	0.18%	0.14%
Aesthetics	Literature-Short Stories	52856	6.78%	5.21%
Aesthetics	Literature-Speeches	361	0.05%	0.04%
Aesthetics	Literature-Text Books (School)	8272	1.06%	0.82%
Aesthetics	Literature-Travelogues	26059	3.34%	2.57%
Commerce	Business	2064	71.79%	0.20%
Commerce	Career and Employment	651	22.64%	0.06%
Commerce	Tourism	160	5.57%	0.02%
Mass Media	Business News	427	0.40%	0.04%
Mass Media	Discussions	3375	3.17%	0.33%
Mass Media	Editorial	1505	1.41%	0.15%
Mass Media	General News	23121	21.72%	2.28%
Mass Media	Health	7925	7.44%	0.78%
Mass Media	Interviews	1060	1.00%	0.10%
Mass Media	Political	57737	54.23%	5.70%
Mass Media	Religious/Spiritual News	750	0.70%	0.07%

Category	Sub Category	Word Count	Percentage (within Sub domain)	Overall Percentage
Mass Media	Social	2120	1.99%	0.21%
Mass Media	Sports News	8207	7.71%	0.81%
Mass Media	Weather	244	0.23%	0.02%
Official Document	Administration	525	9.35%	0.05%
Official Document	Legislature	273	4.86%	0.03%
Official Document	Police Documents	4815	85.78%	0.48%
Science and Technology	Astronomy	103	0.34%	0.01%
Science and Technology	Biology	4007	13.17%	0.40%
Science and Technology	Biotechnology	388	1.28%	0.04%
Science and Technology	Botany	12623	41.49%	1.25%
Science and Technology	Chemistry	667	2.19%	0.07%
Science and Technology	Engineering-Electronics Communication	194	0.64%	0.02%
Science and Technology	Engineering-Others	652	2.14%	0.06%
Science and Technology	Environmental Science	2041	6.71%	0.20%
Science and Technology	Geology	1865	6.13%	0.18%
Science and Technology	Medicine	4390	14.43%	0.43%
Science and Technology	Physics	2290	7.53%	0.23%
Science and Technology	Zoology	1201	3.95%	0.12%
Social Sciences	Demography	8660	9.79%	0.85%
Social Sciences	Economics	995	1.13%	0.10%
Social Sciences	Education	1336	1.51%	0.13%
Social Sciences	Food and Wellness	3733	4.22%	0.37%
Social Sciences	Geography	1388	1.57%	0.14%
Social Sciences	Health and Family Welfare	802	0.91%	0.08%
Social Sciences	History	8221	9.30%	0.81%
Social Sciences	Linguistics	9608	10.87%	0.95%
Social Sciences	Personality Development	769	0.87%	0.08%

Category	Sub Category	Word Count	Percentage (within Sub domain)	Overall Percentage
Social Sciences	Philosophy	12583	14.23%	1.24%
Social Sciences	Political Science	596	0.67%	0.06%
Social Sciences	Religion/Spiritual	36691	41.49%	3.62%
Social Sciences	Sociology	3045	3.44%	0.30%

Table 2: Representation of Sub domains in Kashmiri Raw Text Corpus Vol. II

1.3 REFERENCES

1. Ramamoorthy, L., Narayan Choudhary & Shahid Mushtaq Bhat. 2019. A Gold Standard Kashmiri Raw Text Corpus. Central Institute of Indian Languages, Mysore.
2. Choudhary, Narayan & L. Ramamoorthy. 2019. "LDC-IL Raw Text Corpora: An Overview" in Linguistic Resources for AI/NLP in Indian Languages, Central Institute of Indian Languages, Mysore. pp. 1-10.
3. Bi Bi Mariyam et.al. 2019. "Kashmiri Raw Text Corpus" in Linguistic Resources for AI/NLP in Indian Languages, Central Institute of Indian Languages, Mysore. pp. 61-64.