LDC-IL
Corpus Insights

Editors:
Dr. Rejitha K. S.
Dr. Narayan Kumar Choudhary

LDC-IL CORPUS INSIGHTS

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.

Editors:
Dr. Rejitha K.S.
Dr. Narayan Kumar Choudhary

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

LDC-IL Corpus Insights

Editors: Dr. Rejitha K. S., Dr. Narayan Kumar Choudhary

e-ISBN: 978-93-48633-33-0
CIIL Publication No.: 1516

First published: AD 2025 March :: *Phalguna 1946 Śaka*

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: Rajesha N.

TABLE OF CONTENTS

FIGURES

TABLES

1 THE MOTHER TONGUE PARALLEL TEXT CORPUS OF INDIA

Narayan Kumar Choudhary, Rejitha K .S.

1.1 PARALLEL CORPUS

A parallel corpus consists of an original text in one language along with its equivalent translations into one or more other languages. The most basic form of a parallel corpus involves only two languages, where one corpus is an exact translation of the other. However, some parallel corpora encompass multiple languages. Moreover, the direction of translation does not necessarily remain consistent. When multiple languages are involved, their inter-relationships can vary in nature.

In parallel corpora, the target language texts often exhibit influence from the source language. This influence manifests as deviations in translation that are induced by the structure and characteristics of the source language. However, these deviations are not typically classified as errors in the conventional sense. Instead, they reflect specific lexical and syntactic choices of the target language or target text. Consequently, such target texts may be considered unrepresentative of the natural linguistic patterns of the target language.

A parallel corpus serves as a fundamental resource across various linguistic and computational fields. It plays a crucial role in lexicography, translation systems and information retrieval, making it an essential tool for cross-linguistic studies. In grammatical and lexicographical research, parallel corpora facilitate contrastive and typological analysis by providing direct comparisons between languages. It supports knowledge acquisition for machine translation in Natural Language Processing tasks, which are base of the development of any automated statistical translation systems. In language learning and teaching, parallel corpora offer authentic contrastive language data, aiding learners in understanding linguistic structures and variations across languages. It enables multilingual information retrieval, including cross-language information retrieval, allowing users to access and process information across different languages efficiently.

1.2 LANGUAGES IN INDIA

India is one of the most linguistically diverse countries in the world, shaped by various historical, geographical, and socio-cultural factors. Numerous communities have migrated or invaded India, contributing to its rich linguistic background. India's diverse geographical features have facilitated the development of distinct linguistic varieties. The coexistence of multiple religious and ethnic groups has promoted a multilingual environment, making linguistic diversity a defining characteristic of India.

The languages of India primarily belong to four major language families: Indo-Aryan, Dravidian, Austro-Asiatic, and Tibeto-Burman. According to the 2011 Census, India is home to a total of 121 major languages and 270 mother tongues.

There is no officially declared national language of India. Though the first official language of India is Hindi, English also enjoys the working official language status as per constitution of India. Encouraging linguistic diversity is essential for preserving the country's rich cultural heritage. Nurturing India's multilingualism is crucial for maintaining harmony and inclusivity.

1.3 LDC-IL PARALLEL CORPUS

Despite having 270 tongues as the mother tongues of India as per 2011 census, there is no written evidence of how all of these languages look like. There has been no survey that can give a basic idea about all of these languages. Though several languages of India has a descriptive grammars written on them, most of these 270 mother tongues lack it. Moreover, most of these languages do not have written text readily available and they are mostly unwritten. Following the requirements of the NEP-2020, LDC-IL took upon the task of creating some basic text that could give a peep into the structure of these 270 mother tongues.

With this goal in mind, the LDC-IL planned to develop a parallel corpus on 270 mother tongues of India. It has been stated time an again by various linguists that the whole of South Asia can be defined as one linguistic area. It means that that the languages of India have a lot of linguistic features that they share, despite all the diversities that it displays. Following the principles of language documentation, we started with the goal collecting the common types of sentences and some phrases that the Indian languages typically share.

Thus, LDC-IL framed a structure to accommodate all the linguistic features of languages. The sentences were collected from various example sentences taken from descriptive grammar books on Indian languages, including that of Hindi, Malayalam, Tamil, Santhali, Manipuri and so on. These setences are often part of the language documentation questionnaire as part of the MPI language documentation or the SPPEL language documentation guideline sentences. These sentences were originally conceptualized in the Indian languages (for example, echo-formation is a typical Indian language feature). Following this, their equivalents (with comments wherever necessary) were given in English and Hindi. This became our pivot language language pair.

A total of 5332 sentences were thus collected. They were further categorized and sub-categorized as per their linguistic features. (see the list of list of categories and sub-categories of the linguistic features selected). Given that our goal was to go mother tongues, we started with translating them into the major languages of India, which included all the 22 scheduled languages. This was done to ensure that when we move to mother tongues, the translators could have the choice of getting the sentences in the source language of their choice (as many of the native speakers of these mother tongues are usually conversant with one of these major languages and/or English).

This method helps translators who may have limited proficiency in English. Even though the original source sentences are in English or Hindi or other major languages, the linguistic features are incorporated to ensure that the translators get the essence of the linguistic feature we are trying to capture and return the same feature in their language, if available. By doing so, we have also compromised with the proper structure of native English and went ahead with what would be considered as Indian English, having the Indian linguistic features. Some sentences may have been translated from Hindi or Punjabi to English, leading to the presence of indirect or passive constructions. Such structures can pose challenges in translation, particularly for languages where these constructions are rare or less commonly used or simply absent.

It provides new insights and knowledge about the relationships between languages while also highlighting their differences. These differences and interrelations exist at various linguistic levels, including morphology, lexicon, syntax, and semantics. For instance, morphological features such as gender markers are present in Hindi but absent in Malayalam. Additionally, certain lexical items may not be available in a language if the corresponding concepts or cultural elements do not exist in the language. For example, an equivalent word for "king" is not found in Tripuri language because the concept of a kingdom is unfamiliar to that community. Similarly, indirect questions or positive imperative sentences may lack direct equivalents in some languages.

All syntactic features, including sentence types, compounding, and parts of speech, are comprehensively covered. For example, within the interrogative category, various question forms such as yes-no questions, wh-questions, and tag questions etc. are included in LDC-IL parallel corpus. It is a valuable resource for comparative and cultural studies, as they facilitate the analysis of linguistic structures and cultural influences across languages.

For language technology purposes, the resource may not be big enough to train any statistical of AI models for the given mother tongues. However, with some additional augmentation, a basic model may be created to give the flavour of a particular language/mother tongue of the related language have a higher resource to combine with. Additionally, these datasets may also be useful for various benchmarking purposes in these mother tongues.

1.4 CORPUS FRAMEWORK

The corpus is designed to capture a broad range of structural features across languages, allowing for cross-linguistic comparisons and theoretical analyses. The sentences are framed primarily based on Lingua Descriptive Studies: Questionnaire (Comrie & Smith, 1977), a widely used framework for linguistic description. Additionally, insights have been incorporated from reference works such as A Manual of Linguistic Field Work and Structures of Indian Languages (Abbi, 2001), Language Documentation Handbook (Bhattacharya et al., 2016), and Comprehensive Questionnaire for Tribal Studies (Sivashanmugam & Thayalan, 2012). This diverse set of sources ensures that the questionnaire captures a broad spectrum of linguistic phenomena.

Each language pair of this parallel corpora consists of 5,332 sentences, systematically constructed to capture 159 grammatical categories. These categories include a broad spectrum of linguistic features, including variations in tense, aspect, mood, voice, case, agreement, and syntactic constructions, thereby providing comprehensive coverage of the language's grammatical framework. Proper names within the dataset have been adapted to more familiar Indian names where applicable. The English sentences are derived from Indian language structures, with translations analyzed in the relevant grammatical contexts in each language. This ensures that the dataset remains authentic to the linguistic characteristics of Indian languages.

The linguistic features and corresponding sentence count of the LDC-IL parallel corpus are outlined in Appendix 1.

We hope that this representative corpus will contribute significantly to linguistic research and linguistic, offering a robust foundation for future studies in Indian language processing.

The corpus is developed using an online platform called TranKit. Freelance translators and reviewers are enlisted for each language after successfully passing a proficiency test. The coordination and assignment of translation and review tasks are managed by the LDC-IL staff.

In the first phase of the corpus creation, a total of 32,194 words (of English as source language) have been compiled. Currently, the corpus creation process has been completed for 147 mother tongues.

1.5 SUMMARY OF THE MOTHER TONGUE PARALLEL TEXT CORPUS OF INDIA

The Mother Tongue Parallel Text Corpus of India Vol.I comprising English and 147 mother tongues of India. Each corpus comprises of a total of 5,332 sentences/phrases, systematically structured based on 159 grammatical categories. Appendix 2 gives the word count and character count of each languages. Appendix 3 gives the details of the 413 translators/reviewers who have contributed to this work. The total word count of the corpus across all 147 mother tongues is 4404845 (i.e more than 4.4 million tokens) and the total character count is 23374289 (i.e. 23.3 million).

1.6 REFERENCE

1. Comrie, Bernard & Smith, Norval. 1977. Lingua Descriptive Studies: Questionnaire. Amsterdam: North-Holland Publishing Company.
2. Abbi, Anvita. 2001. A Manual of Linguistic Field Work and Structures of Indian Languages. Muenchen: Lincom Europa.
3. Bhattacharya, Krishna et al. (Eds.) 2016. Language Documentation Handbook. Mysore: CIIL.
4. Sivashanmugam, C. & Thayalan, V. 2012. (Comp.) Comprehensive Questionnaire for Tribal Studies. Coimbatore: Dept. of Linguistics, Bharathiar University.

1.7 Appendix 1: Grammatical Features Covered

Sl. No.	Feature Category	Feature Sub-categry	Count
1	Speech type	Direct Speech	28
2	Speech type	Indirect Speech	43
3	Interrogative	YES-NO Question	54
4	Interrogative	Wh-Question	135
5	Statement	Declarative Sentence	58
6	Interrogative	Tag question	35
7	Interrogative	Alternative question	28
8	Interrogative	Indirect Question	9
9	Interrogative	Constituent of the main clause questioned	37
10	Interrogative	Constituent of subordinate clause questioned	65
11	Interrogative	Noun Clause and reported speech	12
12	Interrogative	Constituent of noun phrase that can be questioned	45
13	Interrogative	Element of Postpositional Phrase can be questioned	19
14	Interrogative	Element of coordinating structures that can be questionned	31
15	Interrogative	Element within coordinated structure can be questionned	8
16	Interrogative	Element of sentence constituent that can be questioned.	33
17	Interrogative	Element of the sentence that can be questioned as echo- question	7
18	Interrogative	Yes-no echo question	46
19	Interrogative	Question word echo Question	42
20	Interrogative	Response particle 'yes' and 'no'	7
21	Imperative Sentence	Positive imperative form	102
22	Imperative Sentence	Negative imperative form	33
23	Statement	Negative sentence	11
24	Imperative Sentence	Person/Number Combination	19
25	Statement	Simple Sentence	7
26	Imperative Sentence	Interrogative as request	7
27	Subordination	Noun Clause	160
28	Subordination	Adjectival Clause (Relative Clause)	256
29	Subordination	Sub-Adverbial Clause	171
30	Structural Question	Internal Structure of the sentence	251
31	Structural Question	Adjectival Phrase	25
32	Structural Question	Adverbial Phrase	25
33	Structural Question	Postpositional Phrase	39
34	Structural Question	Arguments of Postposition	5
35	Structural Question	Noun Phrase	79
36	Coordination	General expression	63

37	Coordination	Coordination- and	52
38	Coordination	Coordination- but	13
39	Coordination	Coordination -or	33
40	Coordination	Coordination and accompaniment	98
41	Negation	Sentence negation	42
42	Negation	Negation of Verbal sentence	9
43	Negation	Constituent Negation	28
44	Negation	Multiple negatives	29
45	Negation	Universal Negation	73
46	Anaphora	Means of Expressing anaphora	38
47	Anaphora	Personal Pronoun	12
48	Anaphora	Reflexive Pronoun	14
49	Anaphora	Reciprocal Pronoun	7
50	Anaphora	Domains of anaphora	13
51	Reflexive	Means of expressing reflexivity-invariable form	52
52	Reflexive	Verbal affix	7
53	Reflexive	Position of the reflexive pronoun	4
54	Reciprocal	Means of expression	51
55	Reciprocal	Superlative	4
56	Comparison	Correlative comparison	19
57	Comparison	Superlative	5
58	Comparison	Comparison	1
59	Equative	Comparison	1
60	Equative	Means of expression	39
61	Posession	Possessive Noun Phrase	48
62	Posession	Possessive Noun	29
63	Emphasis	Non-contradictory emphasis	20
64	Emphasis	Contradictory emphasis	10
65	Emphasis	Particle	15
66	Emphasis	Constituent emphasis	15
67	Emphasis	Movement of emphasised elememt	24
68	Emphasis	Clefting	24
69	Emphasis	YES-NO Question	4
70	Statement	Simple Sentence	3
71	Emphasis	Simple Sentence	5
72	Emphasis	Reflexive Pronoun	1
73	Emphasis	Exclamatory	1
74	Subordination	Sub-Subordinate clause to Main clause	1
75	Heavy Shift and other Movement Rule	Subordinate clause to Main clause	45
76	Topic	General expression	1
77	Interrogative	Wh-Question	1
78	Negation	General expression	1

79	Minor sentence type	Greetings	17
80	Minor sentence type	Exclamatory	11
81	Minor sentence type	Brief answer and response	7
82	Minor sentence type	Nonfinite verb in question	2
83	Minor sentence type	Sentence Adverbial	5
84	Minor sentence type	other	7
85	Sentence Type	Simple Sentence	1
86	Inflectional Morphology	Nominative Case	3
87	Inflectional Morphology	Accusative Case	11
88	Inflectional Morphology	Dative Case	7
89	Inflectional Morphology	Benefactive Case	2
90	Inflectional Morphology	Locative Case	29
91	Inflectional Morphology	Movement of emphasised elememt	2
92	Inflectional Morphology	Sociative Case	3
93	Inflectional Morphology	Simple Sentence	3
94	Inflectional Morphology	Ablative Case	6
95	Inflectional Morphology	Instrumental Case	5
96	Inflectional Morphology	Possessive Case	1
97	Inflectional Morphology	Expression of syntactic function- Intransitive or Transitive Verb	65
98	Inflectional Morphology	Expression of syntactic function- Subject of Copular Verb	22
99	Inflectional Morphology	Expression of syntactic function - Direct object	28
100	Inflectional Morphology	Expression of syntactic function - Indirect object	19
101	Inflectional Morphology	Expression of syntactic function Object governed by verb	16
102	Inflectional Morphology	Expression of syntactic function-Subject Compliment	7
103	Inflectional Morphology	Expression of syntactic function-Object Compliment	5
104	Inflectional Morphology	Nonlocal semantic function	310
105	Inflectional Morphology	Local semantic function	242
106	Inflectional Morphology	Location in time	162
107	Inflectional Morphology	Number Marking system	7
108	Inflectional Morphology	Noun class	12
109	Inflectional Morphology	Noun Phrase-Definiteness	17
110	Inflectional Morphology	Noun Phrase-Indefiniteness	17
111	Inflectional Morphology	Noun Phrase-Referential and non-referential Indefiniteness	11
112	Inflectional Morphology	Noun Phrase-Genericness	12
113	Pronoun	Pronoun-Personal Pronoun	23
114	Pronoun	Pronoun-Indefinite pronoun	12
115	Pronoun	Pronoun-Emphatic pronoun	2
116	Pronoun	Pronoun-Reflexive Pronoun	18
117	Pronoun	Pronoun-Reciprocal Pronoun	6
118	Pronoun	Pronoun-Interrogative Pronoun	7

119	Pronoun	Pronoun-Relative Clause	1
120	Pronoun	Pronoun-Demonstration	22
121	Verb - Inflection	Voice	108
122	Verb - Inflection	Present tense	78
123	Verb - Inflection	Past tense	16
124	Verb - Inflection	Future tense	38
125	Verb - Inflection	Aspect	300
126	Verb - Inflection	Mood	226
127	Verb - Inflection	Causative	57
128	Verb - Inflection	Defective verb	8
129	Verb - Inflection	Finite verb	8
130	Verb - Inflection	Nonfinite verb	7
131	Verb - Inflection	Lexical verb as auxiliary	2
132	Verb - Inflection	Verb Agreement-Person number gender	37
133	Adjective	Adjective-Attributive Adjective	33
134	Adjective	Adjective-Predicative Adjective	42
135	Adjective	Adjective-Comparison	39
136	Adverb	Adverb-Comparison	16
137	Adverb	Adverb-Quality	14
138	Adverb	Adverb-General expression	30
139	Adposition	Adposition-Postposition	64
140	Numeral -Inflection	NumInflect-Ordinal with Nominal suffix	4
141	Numeral -Inflection	NumInflect-Numeral with Personal Noun	10
142	Numeral -Inflection	NumInflect-Quantifier	60
143	Derivational Morphology	Prefixing	7
144	Derivational Morphology	Derived Noun	9
145	Derivational Morphology	Noun from verb	2
146	Derivational Morphology	Nominal marked for tense	5
147	Derivational Morphology	Deverbal Noun	5
148	Derivational Morphology	Noun from Adjective	5
149	Derivational Morphology	Derived verb	5
150	Derivational Morphology	Derived Adjective	8
151	Derivational Morphology	Derived Adverb	2
152	Compounding	Noun noun compound	3
153	Compounding	Part of	2
154	Compounding	Time and season	3
155	Compounding	Coordinate compounds	3
156	Compounding	Echo compounds	2
157	Compounding	Complex compound.	4
158	Compounding	Verb compound	14
159	Compounding	Conjunct Verb	4

Table 1: Grammatical Features Covered

1.8 APPENDIX 2: LANGUAGES INVOLVED AND VOLUME

Sl. No.	Language Name	Word Count	Character Count
1	English	32194	158680
2	Anal	30783	161801
3	Angami	28527	150397
4	Apatani	33911	167364
5	Are	22222	135159
6	Assamese	26055	147249
7	Awadhi	32081	141965
8	Bagheli/Baghel Khandi	32164	142480
9	Bagri Rajasthani	32573	137947
10	Bagri	32046	136098
11	Balti	32919	163480
12	Bengali	26950	146668
13	Bhadrawahi	29971	135769
14	Bharmauri/Gaddi	32351	150257
15	Bhojpuri	31893	141872
16	Bilaspuri Kahluri	32915	148180
17	Bodo	24560	166299
18	Brajbhasha	32625	143975
19	Bundeli/Bundel khandi	32595	140228
20	Chakru/Chokri	28640	130608
21	Chambeali/Chamrali	31853	155017
22	Chang	27641	160636
23	Chhattisgarhi	31973	139986
24	Chirr	30390	172997
25	Chungli	27987	165226
26	Churahi	29579	144939
27	Coorgi/Kodagu	23604	156040
28	Deori	24529	158069
29	Dhundhari	32067	145546
30	Dimasa	23681	152797
31	Dogri	31934	147576
32	Gangte	39306	176058
33	Garhwali	29485	135496
34	Garo	24482	178029
35	Gojri/Gujjari/Gujar	29139	142950
36	Gujarati	29060	144914
37	Gujari	33476	140148
38	Halabi	31696	151964
39	Handuri	32831	146855
40	Hara/Harauti	34537	155620

41	Haryanvi	32885	138071
42	Hindi Multani	32136	165425
43	Hindi	32905	147098
44	Irula/Irular Mozhi	22909	164665
45	Kabui	32841	198758
46	Kachchhi	30410	143091
47	Kangri	31135	154850
48	Kannada	21972	167366
49	Karbi/Mikir	30400	183861
50	Kashmiri	30820	163222
51	Khandeshi	27088	153178
52	Khari Boli	31931	142150
53	Khasi	41762	184587
54	Khezha	29611	150008
55	Khiemnungan	30359	157110
56	Khortha/Khotta	31131	141319
57	Kisan	26035	179192
58	Kodava	25044	166544
59	Kokbarak	27187	175194
60	Kolami	23108	153502
61	Koli	27388	147969
62	Kom	34842	193818
63	Konda	24549	161416
64	Konkani	25384	144901
65	Konyak	34232	190414
66	Koya	23035	162504
67	Kudubi/Kudumbi	24371	148209
68	Kuki	32704	174724
69	Kurmali Thar	28825	148514
70	Ladakhi	5611	155372
71	Lepcha	35379	182292
72	Liangmei	30096	168701
73	Limbu	23296	164517
74	Lotha	31952	171180
75	Lyngngam	36398	172650
76	Magadhi/Magahi	32616	149260
77	Maithili	30959	141773
78	Malayalam	20955	168051
79	Malvi	32454	142183
80	Manipuri	23963	146608
81	Mao	28802	163112
82	Mara	35083	171595
83	Maram	31139	181367

84	Marathi	26327	148721
85	Maring	25073	160556
86	Mech/Mechhia	24707	163754
87	Mewari	31252	139194
88	Mewati	31268	144110
89	Miri/Mising	23024	162508
90	Mishmi	37200	191465
91	Mizo	37596	170734
92	Mongsen	31847	161278
93	Monpa	30510	164566
94	Mundari	31470	166724
95	Muwasi	30762	148990
96	Nawait	29110	145844
97	Nepali	27397	155604
98	Nimadi	32253	137392
99	Nissi	32368	149388
100	Nocte	36634	186628
101	Odia	26141	147972
102	Pahari	32270	139064
103	Paite	32648	152920
104	Palmuha	31398	136462
105	Pania	21731	163599
106	Pawari/Powari	31903	140675
107	Phom	26936	159549
108	Pnar/Synteng	41968	173704
109	Pochury	31149	166904
110	Poula	30306	153054
111	Punjabi	33794	155362
112	Purkhi	31518	168754
113	Rai	25213	173937
114	Rajasthani	31096	143551
115	Reang	36337	164934
116	Rengma	33886	170815
117	Rongmei	37776	200564
118	Sadan/Sadri	31676	144910
119	Sambalpuri	27783	138676
120	Sangtam	28716	173970
121	Sanskrit	24136	152626
122	Santali	32902	188001
123	Saurashtra/Saurashtri	22666	153600
124	Sema	30803	167974
125	Shina	27916	135135
126	Sindhi	33053	156540

127	Sirmauri	30358	144588
128	Sugali	22924	131433
129	Surjapuri	30399	141982
130	Talgalo	33074	151095
131	Tamil	22965	192135
132	Tangkhul	28405	194103
133	Telugu	22829	159641
134	Thado	33759	172320
135	Tibetan	5497	176021
136	Tikhir	29964	172906
137	Tripuri	32631	162746
138	Tulu	23245	150029
139	Urdu	34329	147375
140	Vaiphei	39424	171508
141	Wagdi	31752	130757
142	Wancho	34611	167197
143	Yerava	24453	153083
144	Yerukala/Yerukula	22375	148657
145	Yimchungre	28514	183258
146	Zeliang	33883	160654
147	Zemi	36531	170892
148	Zou	35477	173541

Table 2: Languages Involved And Volume

1.9 APPENDIX 3: TRANSLATORS AND REVIEWERS

Sl. No.	Language	Translators / Reviewers
1	Assamese	Angelina B. Dihingia
		Priyanshe Adhyapak
		Bikash Chetia
		Puja Das
		Udipta Saikia
2	Bengali	Arpita Poddar
		Monalisa Paul
		Agnisha Majumder
		Kuntala Ghosh Dastidar
3	Bodo/Boro	Dr. Bridul Basumatary
		Farson Daimary
		Dr. Bihung Brahma
		Mahananda Brahma
4	Dogri	Kuldeep Kumar
		Rahul Singh
		Diksha Choudhary
		Jatinder Singh
		Suhasi Dhawan
		Skyali Gupta
5	Gujarati	Himani Kanojia
		Natwarlal Modha
		Dr. Rozy Patel
6	Hindi	Neha Dixit
7	Kannada	Sachin S
		Chaitra M N
		Anusha M H
		Rashmi H S
		Nayana H C
		Shauri BP
		Tejaswi G
8	Kashmiri	Zahid Bashir Lone
		Sumeera Bashir
		Sajad Hussain Dar
		Javaid Ahmad Sofi
9	Konkani	Ashalata Navelkar
		Yashwant Gawas
		Siddhi Gawas
		Pooja Tople
		Shweta Parab

		Sunita Gurudas Kanekar
10	Maithili	Sharda Jha
		Dharmbir
11	Malayalam	Dr. Resmi K. S.
		Anoop G. P
		Arya Navya A V
12	Manipuri	Dr. Nameirakpam Amit
		Thaodem Romen Meitei
		Dr. Bachaspatimayum Premabati Devi
		Dr. Thounaojam Harimohon Singh
13	Marathi	Neha Satish Bandekar
		Arvind Ashok Tangadi
		Pratibha Jitendra Dongare
		Apurva Arvind Sambrekar
14	Nepali	Rupesh Rai
		Padam Chhetri
		Sharmila Sharma
15	Odia	Aditya Nayak
		Parthasarathi Dash
		Debamitra Mishra
16	Punjabi	Rajwinder Kaur
		Sukhvinder Singh
		Manpreet Kaur
		Deepika Rani
17	Sanskrit	Dr. Vinayak Bhat
		Varun R
		Vishwanath M V
18	Santhali	Thakur Prasad Murmu
		Tarapada Soren
		Jayanta Tudu
19	Sindhi	Heena Agnani
		Anita Narayan Kukreja
		Mohit Kukreja
		Mehek Wadhwani
20	Tamil	Dr. Prabakaran P
		Dr. V. Alagumuthu
		Dr. Manimala J
21	Telugu	Dr. Rajarao Dunna
		Ravinder Awgani
		P. Sujatha
22	Urdu	Dr. Syed Majid Ali
		Dr. Md. Farhan
		Dr. Tasneem Zahera Haidry

23	Anal	Seltun heberson anal
		DY Shuthahring Anal
		Daryal Thawar Anal
24	Angami	Rüübino Peseyie
		Thepfulenuo Mere
		Vetsino
		Vibeituonuo Mere
25	Apatani	Bamin Lure
		Bamin Yalung
26	Are	Darshini S
		Shruthi B B
		Likhith C Y
27	Awadhi	Dr. Sumedha Shukla
		Vinod Kumar Tiwari
28	Bagheli/Baghel Khandi	Abhyuday Pratap Singh
		Alankrita Singh
		Priyanshu Kushwaha
29	Bagri	Sukhvinder Kaur
		Babbu Kaur
30	Bagri Rajasthani	Dr. Mohan Lal
		Dr.Surinder Kumar
		Gurleen Kaur
		Ranjeet Kaur
31	Balti	Nazir Hussain
		Anayatullah
32	Bhadrawahi	Divya Rathore
		Haresh Kumar
33	Bharmauri/Gaddi	Shubh Karan
		Sunita Devi
34	Bhojpuri	Anupama Tiwari
		Ravi Prakash
		Rajesh Kumar
35	Bilaspuri Kahluri	Ajay Kumar
		Monika Kumari
36	Brajbhasha	Shyam Ratan
		Pramod Rathor
		Omprakash
37	Bundeli/Bundel khandi	Anjana Kishanpuri
		Shivanee
38	Chakru/Chokri	Thukuvelu Sakhamo
		Huvo Thuluo
		Kukhrunelu Theluo
39	Chambeali/Chamrali	Dr. Abdul Kreem

		Dr. Hardeep Singh
		Shahnaz Akhtar
40	Chang	Mosha Mongko
		Narojungla Chang
		Bumo Takum
41	Chhattisgarhi	Jayant Kumar Sahu
		Sanjeev Tiwari
		Dindayal Sahu
42	Chirr	Nutsula
		Yokhumcha
43	Chungli	Tianaro T Lemtur
		Pangernungla
44	Churahi	Hasina
		Fazal Deen
		Mubarak Mohd
45	Coorgi/Kodagu	Thashma K P
		Lavina G S
		Dr. Bhamini Raghavaiah K
46	Deori	Navajyoti Deori
		Lila Kanta Deuri
47	Dhundhari	Ramji Lal Bairwa
		Puran Mal Bairwa
48	Dimasa	Dr. Krithika Barman
		Dr. Monali Longmailai
49	Gangte	T Zoui Gangte
		Paulalson Gangte
		Lamgoulal Gangte
50	Garhwali	Mamta Sharma
		Pradeep Bailwal
		Deepak Bijalwan
51	Garo	Thamalisha W Sangma
		Ringchira G Momin
52	Gojri/Gujjari/Gujar	Naseem Akhtar
		Mushtaq Khalid Chaudary
		Tariq Hussain Abrar
53	Gujari	Azhar Nasir
		Asma Naseer
54	Halabi	Shakuntala Tarar
		Vikram Kumar Soni
55	Handuri	Hari Ram Dhiman
		Vivek Dhiman
56	Hara/Harauti	Kishan Kumar
		Atul Kumar Jain

57	Haryanvi	Dr. Rajinder Kumar
		Ravi kumar
		Kuldeep Gulia
		Ankit Maan
58	Hindi Multani	Dr. Lata Khera
		Gurdatt Singh
59	Irula/Irular Mozhi	Suresh P
		Srinivasan K
60	Kabui	Lanbonlung Longmei
		Kamei Langangmei
61	Kachchhi	Fatima Imtiaz Dhafrani
		Imtiaz Dhafrani
62	Kangri	Ravinder Singh
		Nitish kumar
63	Karbi/Mikir	Khorsing Teron
		Welsing Hanse
64	Khandeshi	Gajanan Suresh Wankhede
		Urmila Patil
65	Khari Boli	Aashi Agarwal
		Vaibhav Singh
		Anshul Garg
66	Khasi	Dr. Kelleney Kitbok Suting
		Dr. P Marlon Brando Rani
		Rikynti L Ryntathiang
67	Khezha	Kewetsou Wetsah
		Koneite U Tsuzu
		Neikhrou Tsuhah
68	Khiemnungan	Songmao M
		Heme
69	Khortha/Khotta	Siddharth Sanket
		Dr. Shilpa
		Dr. Ritu Ghansi
		Basant Kumar
70	Kisan	Nelson Ekka
		Premica Ekka
		Rewalina Ekka
71	Kodava	Revathi
		K M Kusum
		Kuttappa M S
		M V Seetha
		Vidya M D
72	Kokbarak	Jabanika Tripura
		Prasanta Tripura

		Salu Jamatia
73	Kolami	Athram Mothiram
		Athram Rajkumar
		Tekam Tulsiram
74	Koli	Gitanjali Mahesh Pagdhare
		Dr. Vanashri Pradip Phalake
		Oscar Philip Kinny
75	Kom	K Songneihrang Kom
		L Chungneihmun Kom
76	Konda	Janni Somi
		Boyi Gopala Rao
77	Konyak	H Y Nyolong Konyak
		Eswo
78	Koya	Dr. Suryanarayana Kalthi
		Irakam Ramesh
		Payam Srinu
79	Kudubi/Kudumbi	Santhosh Kumar
		Shwetha K
80	Kuki	Lhaineilam
		Hengougin Misao
		Chongpi
		Sehtinmang Kholhou
81	Kurmali Thar	Hardeo Narayan Singh
		Bijraj Mahto
		Gyaneshwar Singh
82	Ladakhi	Lobzang Tsering
		Dr. Neema Tashi
		Dr. Tsering Dolker
83	Lepcha	Churmit Lepcha
		Nimkit Lepcha
84	Liangmei	Lungphubou Abonmai
		Wichamdinbo
85	Limbu	Kausila Subba
		Karnahang Limbu
		Pahang Limboo
86	Lotha	Thungchopeni N Murry
		Yanpomo R Humtsoe
		Nzanmongi Z Ezung
		Dr. Yantsubeni Ngullie
87	Lyngngam	Wandalin S Dkhar
		Balarihun Dkhar
		Resha Nongsiang
		Tyngshainhun Kharsnar

88	Magadhi/Magahi	Dr. Anamika Kumari
		Madhulika
89	Malvi	Parul Upadhyay
		Jaya Rajput
		Prachi Solanki
90	Mao	Ashiihrii Khoziio
		Payia Maheo
		Kholi
91	Mara	N Beithasia
		N Beizatha
		Loisy Khithie
92	Maram	Ph Pungdila Grace
		H Keren
		Lymba Solomon T
93	Maring	Kanshouwa Susie
		Dangsawa Moran Maring
		K Florida Maring
		Ch Abose Moyon
		Shimlung Makunga Maring
94	Mech/Mechhia	Ananta Champromary
		Dr. Bikash Narjinary
95	Mewari	Dr. Reena Menariya
		Dr. Ripudaman Singh Ujjwal
96	Mewati	Hanif Khan
		Ammar Khan
97	Miri/Mising	Dimbeswar Doley
		Bastav Lagachu
		Kalinath Panging
98	Mishmi	Miju Mena
		Tiba Apralo
99	Mizo	Dr. Wendy Lalhminghlui
		Catherine Lalhruaizeli
		Vanlalhruaizeli Chawngthu
100	Mongsen	Amongkumba I kichu
		Imchalemla Longchar
101	Monpa	Nima Dorjee
		Pema chotton
102	Mundari	Dev Kumar Munda
		Amar Topno
		Hercules Singh Munda
103	Muwasi	Arjun
		Sachin
104	Nawait	Mohammad Zubair Jukaku

		Mohammed Rayees
105	Nimadi	Nikita
		Ajay Patel
		Shivpal kalam
		Rohit Patel
106	Nissi	Taba Yami
		Likha Bai
		Nabam Tai Hina
107	Nocte	Dr. Doli Tesia
		Dr. Dancha Tongluk
		Dr. Chalit Sumnyan
108	Pahari	Syed Murtuza Hussain
		Dr. Syed Mahir Hussain Jafri
109	Paite	Ching
		Ngaithianhoih
		Nempalching Hangshing
		Eddie Hnunrousiam Valte
110	Palmuha	Parth Ranjan
		Shreekant Dubey
111	Pania	Nandakumar S
		Anjali Bhaskaran
112	Pawari/Powari	Dr. Tufan singh pardhi
		Dr. Shobha
113	Phom	Rebecca Wandoi Phom
		V Shau-au
		Lily Phom
114	Pnar/Synteng	Egira Shadap
		Juhhi Nanghuloo
		Lammidaka Pohleng
115	Pochury	Chukhutho Nyusou
		Lojirho
		Lolia Ngouri
		Wuojisie Nyusou
116	Poula	L V Vaveiru
		K Dosou
		R S Japhet Khailunii
117	Purkhi	Sajjad Ali
		Mohd Issa
118	Rai	Taraman Rai
		Dil Kumar Rai
		Bir Bahadur Rai
		Pratima Rai
119	Rajasthani	Dr. Madan Gopal Ladha

		Dr. Neeraj Daiya
		Dr. Sharada Krishna
120	Reang	Rusan Reang
		Rotnojoy Reang
121	Rengma	Sinyhunlo Kent
		Akenye Seb
		Kenyuni Kent
122	Rongmei	K Gloria Phaomei
		Elisha Panmei
123	Sadan/Sadri	Radheshyam Lakra
		Bibha Rani Topno
124	Sambalpuri	Suman Rani Panda
		Astha Ayusha Mall
125	Sangtam	Salome Sangtam
		Sethrila Thongtsar
126	Saurashtra/Saurashtri	K. S. Senthilkumar
		K S Malini
127	Sema	Pikali L Assumi
		Honikali Lohe
128	Shina	Mohd Shafi
		Mukhtar Ahmad
129	Sirmauri	Archna Kumari
		Nantram
130	Sugali	Dr. Prasad Naik Mude
		Bukke santhoshnaik
131	Surjapuri	Sanowar jahan
		Zoofishan khanam
		Khushnuma Begum
132	Talgalo (galo)	Nivam Rekhung
		Jirken Bogo
133	Tangkhul	Dr. Yatmi Luikham
		Dr. Maireiwon Ningshen
		Rinsomi Luikham
		Somiwon A Shishak
134	Thado/Thadou	Dr. Mangvung Hemminlan Haokip
		Dr. Zamminlien
		Mongngaichong Touthang
		Boishi Neijalhei khongsai
135	Tibetan	Dr. Lhamo Tso
		Lobsang Gendun
136	Tikhir	Shiutsu K Thongliu
		Longtsu k lams
137	Tripuri	Kwrwng Tripura

		Salka Tripura
		Sushil Tripura
		Bhabesh Tripura
138	Tulu	Sai Geeetha
		Yogitha Shetty
		Soumya Rao
		Dr. Rajashree
		Pradyoth Hegde
139	Vaiphei	P Damlalson Vaiphei
		Genthianson Vaiphei
		P Felboi Vaiphei
140	Wagdi	Hoshang Panchal
		Kushagra Panchal
141	Wancho	Lemnon Wangjen
		Omwang wangjen
142	Yerava	Poovaiah M P
		Accamma M S
143	Yerukala/Yerukula	E Mahalaxmi
		S Pravallika
144	Yimchungre	Nagayimla Yimchunger
		Hantsula
145	Zeliang	Lungchiepyile
		Machipeung Thou
		Kangzangyile Khate
146	Zemi	Atule Heikha
		Ichilungle
		Iraicule
		Haidobabe Hingleu
147	Zou	M Joseph Khamgoulian Zou
		Thangsianlal William

Table 3: Translators And Reviewers

2 A GOLD STANDARD RAJASTHANI RAW TEXT CORPUS

Ankita Tiwari, Narayan Kumar Choudhary

2.1 INTRODUCTION

The Linguistic Data Consortium for Indian Languages (LDC-IL) is actively collecting written materials from various mother tongues. What began with Chhattisgarhi has now extended to Rajasthani. This continuous effort highlights LDC-IL's dedication and commitment to preserving and promoting mother tongues; and paving the way for their advancement in language technology.

The LDC-IL has chosen to release this dataset as the Rajasthani Raw Text Corpus, encompassing its recognized linguistic varieties, which are elaborated upon in the subsequent paragraph. Although the Rajasthani is listed as one of the mother tongues in the 2011 Census, it also appears to serve as a broad term encompassing the diverse languages spoken in Rajasthan, which may be collectively referred to as Rajasthani. This corpus aims to capture its linguistic diversity comprehensively.

A corpus effectively captures the linguistic nuances and unique characteristics of a language when it meets two essential criteria: a significant volume and an authentic representation across various domains. A language's text corpus serves as a valuable resource for scientific exploration, providing reliable evidence of its features and enabling a deeper understanding of its distinct properties.

The Rajasthan state hosts a wide array of linguistic forms that hold deep cultural and historical significance. More than just a means of communication, Rajasthani serves as a vital medium for preserving and transmitting the rich heritage, traditions, and folklore of Rajasthan. Rajasthani is recognized by the Government of India as a Western Indo-Aryan variant of Hindi, which is predominantly spoken within Rajasthan. According to the Census of India, 2011 approximately 51,968,635 individuals across India identified Rajasthani (including all the varieties) as their spoken language.[1] Rajasthani may be considered an umbrella term which includes a diverse linguistic group encompassing multiple mother tongues spoken across various regions of Rajasthan. A list of mother tongues with ISO codes listed below in the Table 3.

#	Mother Tongue	ISO 639-3 Language Code	Number of Speakers (Across India)
1	Rajasthani	(ISO 639-3: raj)	2,58,06,344
2	Marwari	(ISO 639-3: rwr)	78,31,749
3	Mewari	(ISO 639-3: mtr)	42,12,262
4	Wagdi	(ISO 639-3: wbr)	33,93,991
5	Hara/Harauti	(ISO 639-3: hoj)	29,44,356
6	Dhundhari	(ISO 639-3: dhd)	14,76,446
7	Bagri Rajasthani	(ISO 639-3: bgq)	2,34,227
8	Malvi	(ISO 639-3: mup)	52,12,617
9	Mewati	(ISO 639-3: wtm)	8,56,643

Table 4: ISO Codes for Mother Tongue

[1] https://censusindia.gov.in/nada/index.php/catalog/42458

2.2 Extraction of Textual Materials

For languages with abundant data, LDC-IL selects a few chapters from each book to ensure consistency across domains and linguistic variations within the corpus. However, in the case of Rajasthani, the limited availability of literary text resources necessitated the inclusion of most available materials in their entirety, eliminating the need for data sampling. Nonetheless, selective sampling was applied to certain books containing extensive tables and data, with only the relevant textual content being extracted. This process was carried out with careful consideration to ensure that paragraphs and sentences retained meaningful context.

2.3 Data Collection

Dr. Neeraj Kumar Daiya, a distinguished author from Rajasthan and a Resource Person for LDC-IL, contributed significantly to the collection of Rajasthani textual materials for the development of this corpus while working remotely on a Marwari translation project. He played a pivotal role in facilitating the LDC-IL team's outreach to universities, colleges, publishing houses, and individual authors through emails and phone calls, successfully assisting in the acquisition of PDFs of books and magazines.

2.4 Processing of Raw Data

The books underwent a scanning process where each page's image was assigned a distinct identifier. These images were subsequently uploaded onto the LDC-IL Data portal. The complete dataset was then categorized into multiple tasks based on character count, typically around 30,000 characters per task. These tasks were subsequently imported into the LDC-IL digitization platform, which possesses the capability to perform Optical Character Recognition (OCR) and extract the text content from each file.

2.5 Selection of Language Experts & Digitization Process

Among the candidates who applied for the freelance language expert position at the CIIL, those who met the eligibility criteria were interviewed to assess their proficiency in Rajasthani. Candidates who successfully cleared the initial screening were then given sample tasks to assess their proficiency and accuracy when using the digitization platform. The performance of each candidate was evaluated by an expert on four levels; (1) spelling, (2) grammar (3) punctuation and (4) spacing accuracy. The candidates passing the test were selected for the task of digitization of books. The entire dataset was digitized by the selected candidates using the same tool against monetary remuneration. Once the Digitization process was complete, each task went through a rigorous review process to check the authenticity of the work using the same platform, before finalizing the data.

A dedicated and diverse team of freelance language experts from various regions of Rajasthan, including Mr. Kishan Kumar Dodia, Mr. Satvir, Dr. Neeraj Daiya, Ms. Mamta Kumari, Mr. Puran Mal Bairwa, Mr. Ram Lal Bairwa, and Mr. Hoshang Panchal, played a crucial role in the

Rajasthani text digitization project. The entire process of digitizing Rajasthani texts and reviewing the digitized materials was carried out remotely by these language experts.

The Rajasthani Raw Text Corpus constitutes a valuable resource for systematically documenting colloquial expressions, idiomatic usage, regional lexicon, and grammatical structures that are fundamental to the development of linguistic processing frameworks. This corpus functions as an extensive repository, preserving the essential linguistic characteristics of Rajasthani textual materials, thereby facilitating scholarly research and computational applications in language processing.

2.6 RAJASTHANI RAW TEXT CORPUS

The total volume of LDC-IL Rajasthani Raw Text Corpus is 11,99,502 words, systematically gathered from 12 books and 62 magazines. The following table gives a summary of the Rajasthani Raw Text Corpus.

#	Category	Sub-Category	Word count	Percentage (Within Sub-Domain)	Overall Percentage
1	Aesthetics	Biographies	1259	0.20%	0.11%
2	Aesthetics	Culture	50628	7.91%	4.23%
3	Aesthetics	Folk Tales	24253	3.79%	2.03%
4	Aesthetics	Literary Texts	201927	31.56%	16.84%
5	Aesthetics	Literature-Children's Literature	7721	1.21%	0.65%
6	Aesthetics	Literature-Criticism	10354	1.62%	0.87%
7	Aesthetics	Literature-Diaries	18842	2.95%	1.58%
8	Aesthetics	Literature-Letters	237	0.04%	0.02%
9	Aesthetics	Literature-Novels	1773	0.28%	0.15%
10	Aesthetics	Literature-Plays	28171	4.40%	2.35%
11	Aesthetics	Literature-Poetry	908	0.15%	0.08%
12	Aesthetics	Literature-Short Stories	158738	24.78%	13.24%
13	Aesthetics	Literature-Text Books (School)	120591	18.82%	10.06%
14	Aesthetics	Literature-Travelogues	13857	2.17%	1.16%
15	Aesthetics	Mythology	1535	0.24%	0.13%
16	Mass Media	Cinema News	24569	5.49%	2.05%
17	Mass Media	Classifieds	1687	0.38%	0.15%
18	Mass Media	Editorial	305484	68.19%	25.47%
19	Mass Media	General News	41215	9.20%	3.44%
20	Mass Media	Health	1056	0.24%	0.09%
21	Mass Media	Interviews	53944	12.05%	4.50%
22	Mass Media	Letters	16679	3.73%	1.40%
23	Mass Media	Obituary	137	0.04%	0.02%
24	Mass Media	Religious/Spiritual News	1660	0.38%	0.14%
25	Mass Media	Sports News	831	0.19%	0.07%
26	Mass Media	Weather	730	0.17%	0.07%
27	Social Sciences	History	110716	100	9.24%

Table 5: Representation of Sub-Categories in the Rajasthani Raw Text Corpus

The LDC-IL Rajasthani Raw Text Corpus comprises a total of 11,99,502 words. This corpus is classified into three primary domains and further subdivided into 27 distinct sub-categories. The distribution of words across these domains is as follows: the aesthetic domain contains 6,40,794 words, the mass media domain includes 4,47,992 words, and the social sciences domain accounts for 1,10,716 words. Collectively, the corpus consists of 11,99,502 tokens. A comprehensive representation of this distribution is illustrated in the accompanying chart.

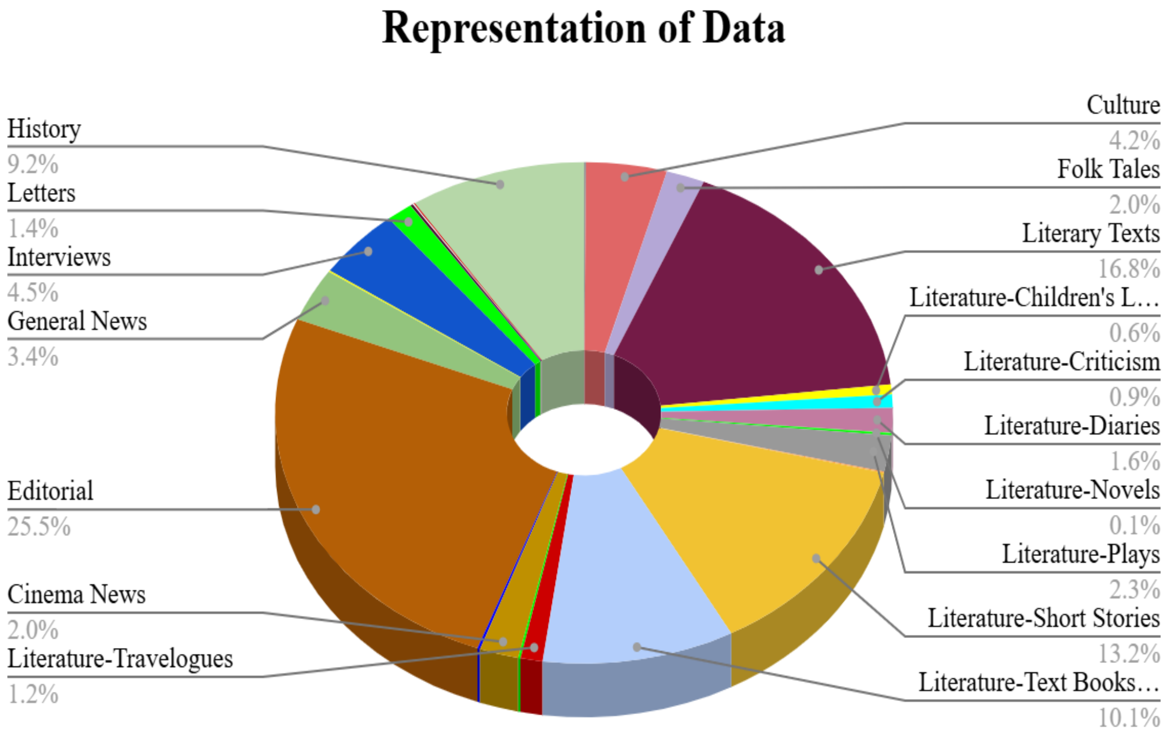

Figure 1: Representation of the Sub-categories in Rajasthani Raw Text Corpus

1.1 References

1. Narayan Choudhary. LDC-IL: The Indian repository of resources for language technology. Lang Resources & Evaluation 55, 855–867 (2021). https://doi.org/10.1007/s10579-020-09523-3
2. Ethnologue. (n.d.). hne|ISO-639-3. Retrieved 09 15, 2023, from SIL: https://iso639-3.sil.org/code/hne
3. Office of the Registrar General & Census Commissioner, India (ORGI). (2022, 04 22). Language. Retrieved from Census of India 2011: https://censusindia.gov.in/nada/index.php/catalog/42458

3 A GOLD STANDARD CHHATTISGARHI RAW TEXT CORPUS VOL. II

Ankita Tiwari, Narayan Kumar Choudhary

3.1 INTRODUCTION

The Linguistic Data Consortium for Indian Languages (LDC-IL) launched an initiative to conduct a fieldwork for the collection of Chhattisgarhi speech data in January 2023. Over time, this initiative expanded to include the compilation of textual resources for building the Chhattisgarhi Raw Text Corpus and released that dataset in January 2024. This initiative has set a significant precedent for other mother tongues as well.

As part of this ongoing initiative, LDC-IL has developed a new Chhattisgarhi dataset containing a substantial number of tokens from multiple domains. Whereas the earlier dataset concentrated largely on the aesthetic domain, the new dataset seeks to provide a more inclusive representation by incorporating data from additional areas, such as news, website and many more, to reduce the prior domain bias. This continuous commitment reflects LDC-IL's commitment and dedication to the preservation and promotion of mother tongues and enhancing the resources to develop language technologies for native languages.

The primary objective of Volume II of the Chhattisgarhi Raw Text Corpus is to collect a vast amount of written material from various domains to preserve the diversity and richness of Chhattisgarhi culture and traditions. And this second volume tries to capture the complexities and unique characteristics of the language by meeting two essential criteria: a substantial volume and an authentic representation of multiple domains.

For languages in which there is ample data, LDCIL selects a few chapters from each book as part of the corpus to maintain uniformity in terms of domains and variety. In the case of Chhattisgarhi, due to the limited availability of literary text resources, all the resources made available to us were selected in entirety to be a part of the text corpus, avoiding the process of data sampling.

Mr. Jayant Sahu, a distinguished author and owner/publisher of the Chhattisgarhi magazine Anjor, alongwith Mr. Sanjeev Tiwari, an advocate and writer, and Mr. Deen Dayal Sahu, a renowned writer from Chhattisgarh, played a significant role in the collection and development of the Chhattisgarhi text corpus (Vol. II) while working remotely as Resource Persons for LDC-IL. Their contributions were noteworthy in compiling Chhattisgarhi textual materials for the corpus. Additionally, they actively facilitated the LDC-IL team's engagement with writers, publishing houses, and individual authors through email and phone communication. Notably, Mr. Tiwari made a remarkable contribution by providing a substantial collection of Chhattisgarhi content published on his website. Their dedicated efforts were crucial in acquiring PDFs of books and magazines, thereby enhancing the corpus development process.

The books underwent a scanning process where each page's image was assigned a distinct identifier. These images were subsequently uploaded onto the LDC-IL Data portal. The complete dataset was then categorized into multiple tasks based on character count, typically around 30,000 characters per task. These tasks were subsequently imported into the LDC-IL digitization platform, which possesses the capability to perform Optical Character Recognition (OCR) and extract the text content from each file.

Among the candidates who applied for the freelance language expert position at CIIL, those who met the eligibility criteria were interviewed to assess their proficiency in Chhattisgarhi. Candidates who successfully cleared the initial screening were then given sample tasks to assess their proficiency and accuracy when using the digitization platform. The performance of each candidate was evaluated by an expert on four levels; 1) spelling, 2) grammar 3) punctuation and 4) spacing accuracy. The candidates passing the test were selected for the task of digitization of books. The entire dataset was digitized by the selected candidates using the same tool against monetary remuneration. Once the Digitization process was complete, each task went through a rigorous review process to check the authenticity of the work using the same platform, before finalizing the data.

A dedicated and diverse team of freelance language experts from different regions of Chhattisgarh, including Mr. Jayant Sahu, Mr. Sanjeev Tiwari, Ms. Hasina, Mr. Sanju Gupta, Mr. Sanjay Gangwal, Mr. Pramod Kumar, Ms. Pushplata, and Mr. Sandeep Dubey played a vital role in the Chhattisgarhi text digitization project. These experts worked remotely to digitize Chhattisgarhi texts and meticulously review the digitized content.

3.2 CHHATTISGARHI RAW TEXT CORPUS VOL. II

The LDC-IL Chhattisgarhi Raw Text Corpus comprises a total of 22,19,592 words, systematically gathered from 33 books, 22 monthly magazines and a website: https://gurturgoth.com/. The following table gives a summary of the Chhattisgarhi Raw Text Corpus.

#	Category	Sub-Category	Word Count	Percentage (Within Sub-Domain)	Overall Percentage
1	Aesthetics	Cinema	149	0.02%	0.01%
2	Aesthetics	Culture	6085	0.57%	0.28%
3	Aesthetics	Folk Tales	57440	5.34%	2.59%
4	Aesthetics	Literary Texts	94202	8.76%	4.25%
5	Aesthetics	Literature-Children's Literature	12893	1.20%	0.59%
6	Aesthetics	Literature-Criticism	36065	3.36%	1.63%
7	Aesthetics	Literature-Novels	115627	10.75%	5.21%
8	Aesthetics	Literature-Plays	16897	1.58%	0.77%
9	Aesthetics	Literature-Short Stories	408728	37.99%	18.42%
10	Aesthetics	Literature-Text Books (School)	7593	0.71%	0.35%
11	Aesthetics	Literature-Travelogues	561	0.06%	0.03%

#	Category	Sub-Category	Word Count	Percentage (Within Sub-Domain)	Overall Percentage
12	Aesthetics	Mythology	319613	29.71%	14.39%
13	Mass Media	Cinema News	4950	0.45%	0.23%
14	Mass Media	Classifieds	85	0.01%	0.01%
15	Mass Media	Discussions	1103	0.09%	0.05%
16	Mass Media	Editorial	97920	8.85%	4.42%
17	Mass Media	General News	931254	84.14%	41.96%
18	Mass Media	Health	29226	2.65%	1.32%
19	Mass Media	Interviews	2502	0.23%	0.12%
20	Mass Media	Letters	1744	0.16%	0.08%
21	Mass Media	Political	2713	0.25%	0.13%
22	Mass Media	Religious/Spiritual News	21931	1.99%	0.99%
23	Mass Media	Social	1360	0.13%	0.07%
24	Mass Media	Sports News	12022	1.09%	0.55%
25	Science and Technology	Agriculture	16585	97.11%	0.75%
26	Science and Technology	Ayurveda	494	2.89%	0.03%
27	Social Sciences	History	1116	5.63%	0.06%
28	Social Sciences	Linguistics	18734	94.38%	0.85%

Table 6: Representation of Sub-categories in Chhattisgarhi Raw Text Corpus Vol. II

This corpus is classified into four primary domains and further subdivided into 28 distinct sub-categories. The distribution of words across these domains is as follows: the aesthetic domain contains 10,75,853 words, the mass media domain includes 11,06,810 words, Science and technology domain includes 17,079 and the social sciences domain accounts for 19,850 words. Collectively, the corpus consists of 22,19,592 tokens. A comprehensive representation of this distribution is illustrated in the accompanying chart.

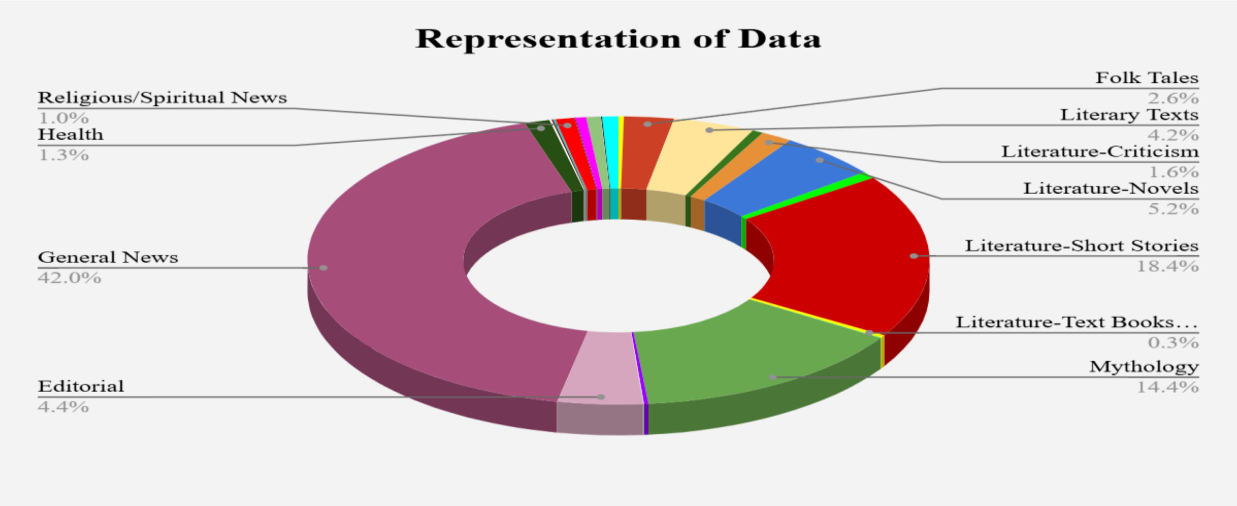

Figure 2: Representation of the Sub-categories in Chhattisgarhi Raw Text Corpus Vol. II

3.3 REFERENCES

4. Narayan Choudhary. LDC-IL: The Indian repository of resources for language technology. Lang Resources & Evaluation 55, 855–867 (2021). https://doi.org/10.1007/s10579-020-09523-3

5. Office of the Registrar General & Census Commissioner, India (ORGI). (2022, 04 22). Language. Retrieved from Census of India 2011: https://censusindia.gov.in/nada/index.php/catalog/42458

6. Ankita Tiwari, S.K. Awasthi , N. Rajesha, G. Manasa, D. Srikanth, N. K. Choudhary, Shailendra Mohan, 2023. A Gold Standard Chhattisgarhi Raw Text Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-19411-64-1: https://data.ldcil.org/chhattisgarhi-raw-text-corpus?tag=Chhattisgarhi

4 A GOLD STANDARD KASHMIRI RAW TEXT CORPUS VOL. II

Dr. Zargar Adil Ahmad

4.1 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.

4.2 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	267140	76.93%
Commerce	177666	0.28%
Mass Media	618096	10.50%
Official Document	406791	0.55%
Science and Technology	727632	3.00%
Social Sciences	936637	8.72%
Total	10,13,658	100%

Table 7: Representation of the Domains in Kashmiri Raw Text Corpus Vol. II

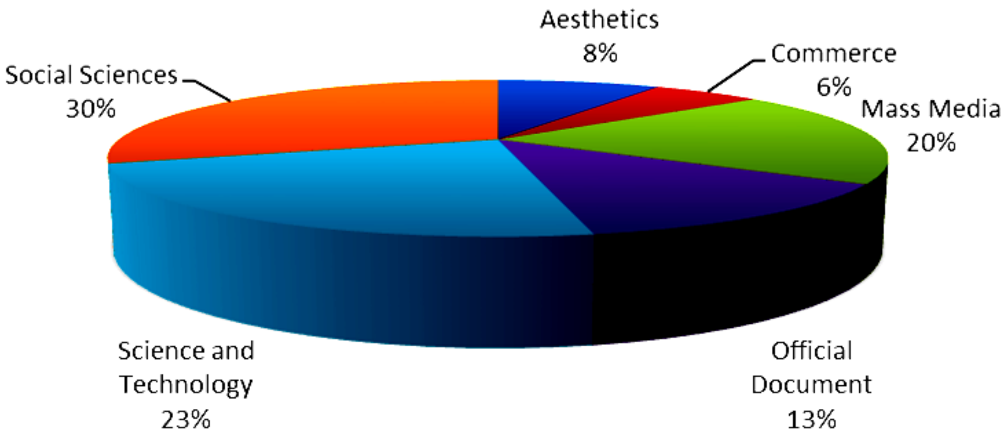

Figure 3: Representation of the Domains in Kashmiri Raw Text Corpus Vol. II

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%

Category	Sub Category	Word Count	Percentage (within Sub domain)	Overall Percentage
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%
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%

Category	Sub Category	Word Count	Percentage (within Sub domain)	Overall Percentage
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%
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 8: Representation of Sub domains in Kashmiri Raw Text Corpus Vol. II

4.3 REFERENCES

7. Ramamoorthy, L., Narayan Choudhary & Shahid Mushtaq Bhat. 2019. A Gold Standard Kashmiri Raw Text Corpus. Central Institute of Indian Languages, Mysore.
8. 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.
9. 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.

5 A GOLD STANDARD MAITHILI RAW TEXT CORPUS VOL. II

Shantanu Kumar, Narayan Kumar Choudhary

5.1 INTRODUCTION

In 2019, the LDC-IL released a Gold Standard Maithili Raw Text Corpus [1], which was developed from available contemporary texts. The corpus creation followed the pre-defined criteria which are mentioned at [2]. The corpus contained 5,316,552 Words and 29,658,942 characters drawn from 499 different titles, including the extracts from Magazine and newspapers that can be referred at [3]. It provides comprehensive representations of major domains such as aesthetics and social sciences, among others. The Maithili 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, the Linguistic Data Consortium for Indian Languages (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.

Maithili is spoken by around 35 million people across India (Census 2011)[2]. As per the linguistic survey made by GA Grierson [4], the Maithili language is spoken widely in the Bihar and Jharkhand states of India. It has several varieties [5]. A few studies mentioned above discuss that the variety of Maithili spoken in the regions of Bhagalpur and the neighboring district is called Angika by the native speakers [6]. The region carries a great cultural and historical significance and is widely used by millions of people across districts. Though the region is primarily claimed as the Angika-speaking region by several works [6] [7] and a few other individual research works, still, based on the government documents and the Census report of 2011[8] by the Govt. of India, the area is considered as the language variety of Maithili. Hence, the area is referred to as the Maithili-speaking region in this document hereafter.

LDCIL selects specific chapters from each book to ensure a balanced representation of domains and variety within the corpus for languages with sufficient data. However, due to the limited availability of literary resources in this particular variety of Maithili, the entire content of all accessible materials was included, avoiding the process of data sampling.

In November 2022, a team of six resource persons- Shantanu Kumar, Rupesh Kumar Pandey, Akanksha Tiwari, Jyoti Kumari, Nikhil Kumar, and Mukesh Kumar was sent to various districts in Bihar and Jharkhand, including Bhagalpur, Banka, Jamui, Munger, Lakhisarai, Khagaria, Godda, and Deoghar, to conduct field work. The team visited villages, universities, colleges, and individual authors across these regions to gather texts in books, magazines, and journals. The team also contacted several authors via email and phone and successfully obtained books through India Post. During the fieldwork of the data collection, there was a huge support by the local authors and individuals in every aspect of the project. There were a few influential literary

[2] https://censusindia.gov.in/nada/index.php/catalog/42458

people who generously contributed directly or indirectly in collecting the text and speech corpus from the remote areas of the locality or from the places wherever possible.

5.2 Maithili Raw Text Corpus Vol. II

A Gold Standard Maithili Raw Text Corpus Vol. II is a comprehensive collection of Maithili language texts, comprising 8,11,680 words from 38 books and 16 magazines. This corpus includes extracts from books 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 Maithili language. The corpus has been meticulously compiled and is available for access through the Linguistic Data Consortium for Indian Languages (LDC-IL). Researchers and developers can utilize this resource to enhance their understanding and applications related to the Maithili language.

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

Domain	Word count	Percentage
Aesthetics	596841	73.53%
Mass Media	184957	22.79%
Social Sciences	29882	3.68%
Total	8,11,680	100%

Table 9: Representation of the Domains in Maithili Raw Text Corpus Vol. II

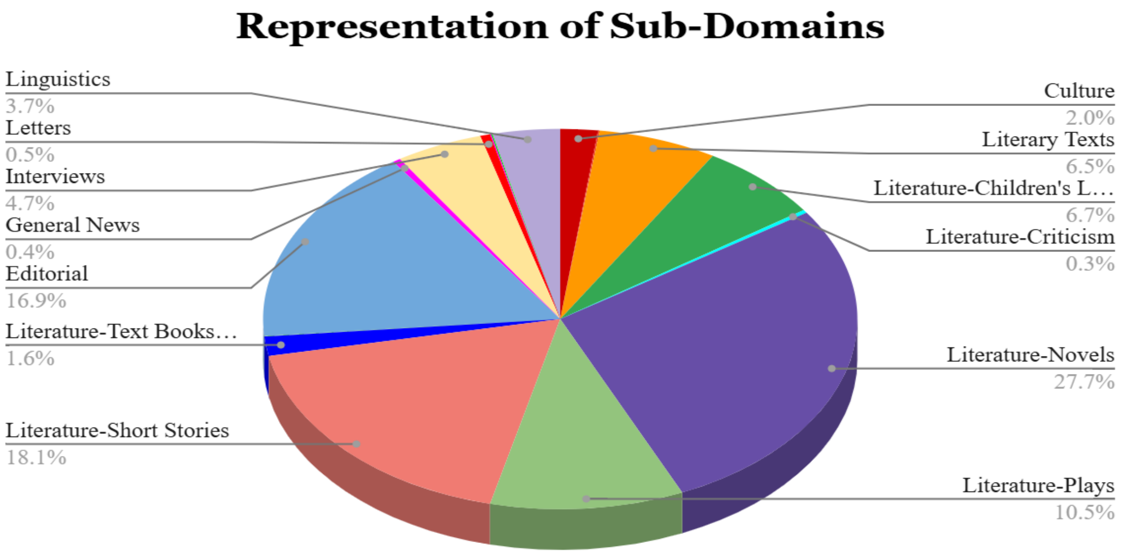

Figure 4: Representation of the Domains in Maithili Raw Text Corpus Vol. II

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	Culture	16,540	2.78%	2.04%
Aesthetics	Folk Tales	608	0.10%	0.08%
Aesthetics	Literary Texts	52,499	8.80%	6.47%
Aesthetics	Literature-Children's Literature	54,426	9.12%	6.71%
Aesthetics	Literature-Criticism	2,192	0.37%	0.28%
Aesthetics	Literature-Novels	224,967	37.70%	27.72%
Aesthetics	Literature-Plays	85,393	14.31%	10.53%
Aesthetics	Literature-Short Stories	146,847	24.61%	18.10%
Aesthetics	Literature-Text Books (School)	13,369	2.24%	1.65%
Mass Media	Cinema News	321	0.18%	0.04%
Mass Media	Editorial	137,438	74.31%	16.94%
Mass Media	General News	3,372	1.83%	0.42%
Mass Media	Interviews	38,446	20.79%	4.74%
Mass Media	Letters	4,339	2.35%	0.54%
Mass Media	Obituary	565	0.31%	0.07%
Mass Media	Social	476	0.26%	0.06%
Social Sciences	Linguistics	29,882	100	3.69%

Table 10: Representation of Sub domains in Maithili Raw Text Corpus Vol. II

5.3 REFERENCES

10. Ramamoorthy, L., Narayan Choudhary & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore.
11. 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.
12. Dinesh Mishra et.al. 2019. "Kashmiri Raw Text Corpus" in Linguistic Resources for AI/NLP in Indian Languages, Central Institute of Indian Languages, Mysore. pp. 73-81.
13. Grierson, GA. 1909, An introduction to the Maithili dialect of the Bihari language as spoken in North Bihar (2 ed.). Calcutta: Asiatic Society of Bengal. pp. xi-xiii.
14. Brass, P. R. (2005). Language, Religion, and Politics in North India. iUniverse, Lincoln, NE
15. Sharma, RM. 2007, Angika Bhasha Ka Dhwanivaigyanik Adhyayan (In Hindi) (II ed.). Shri Tara Geeta Printing Press, Bhagalpur, pp. 130-133.
16. Jha, S. 1958, "The formation of the Maithili language." PhD diss., Luzac London. pp. 5-7.

6 A GOLD STANDARD TELUGU RAW TEXT CORPUS VOL.II

Dr. Modugu Kasimbabu, Narayan Kumar Choudhary

6.1 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, Vishakahppatanam, 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.

6.2 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 11**:** 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 12**:** Representation of the Sub-domains in Telugu Raw Text Corpus Vol. II

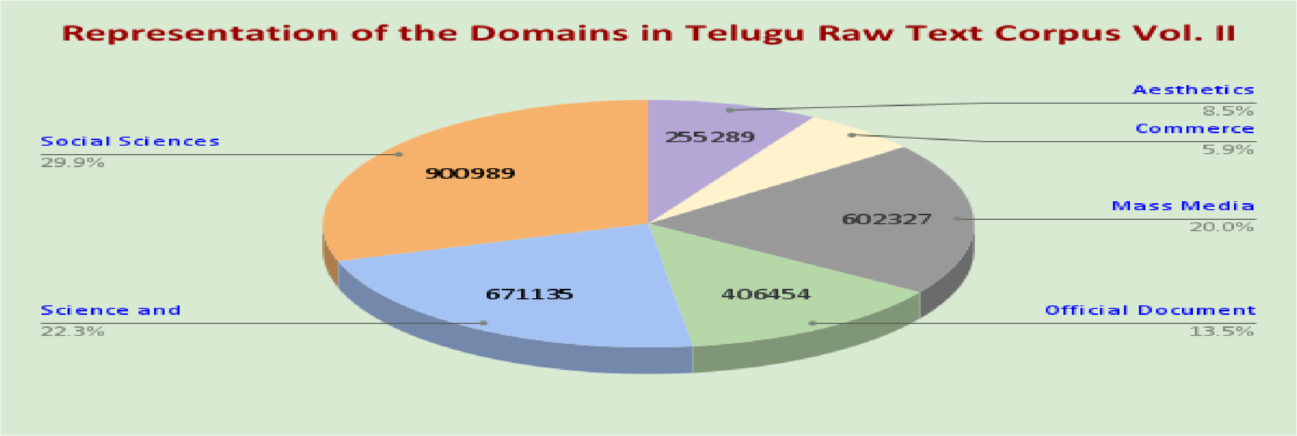

Figure 5: Representation of the Domains in Telugu Raw Text Corpus Vol. II

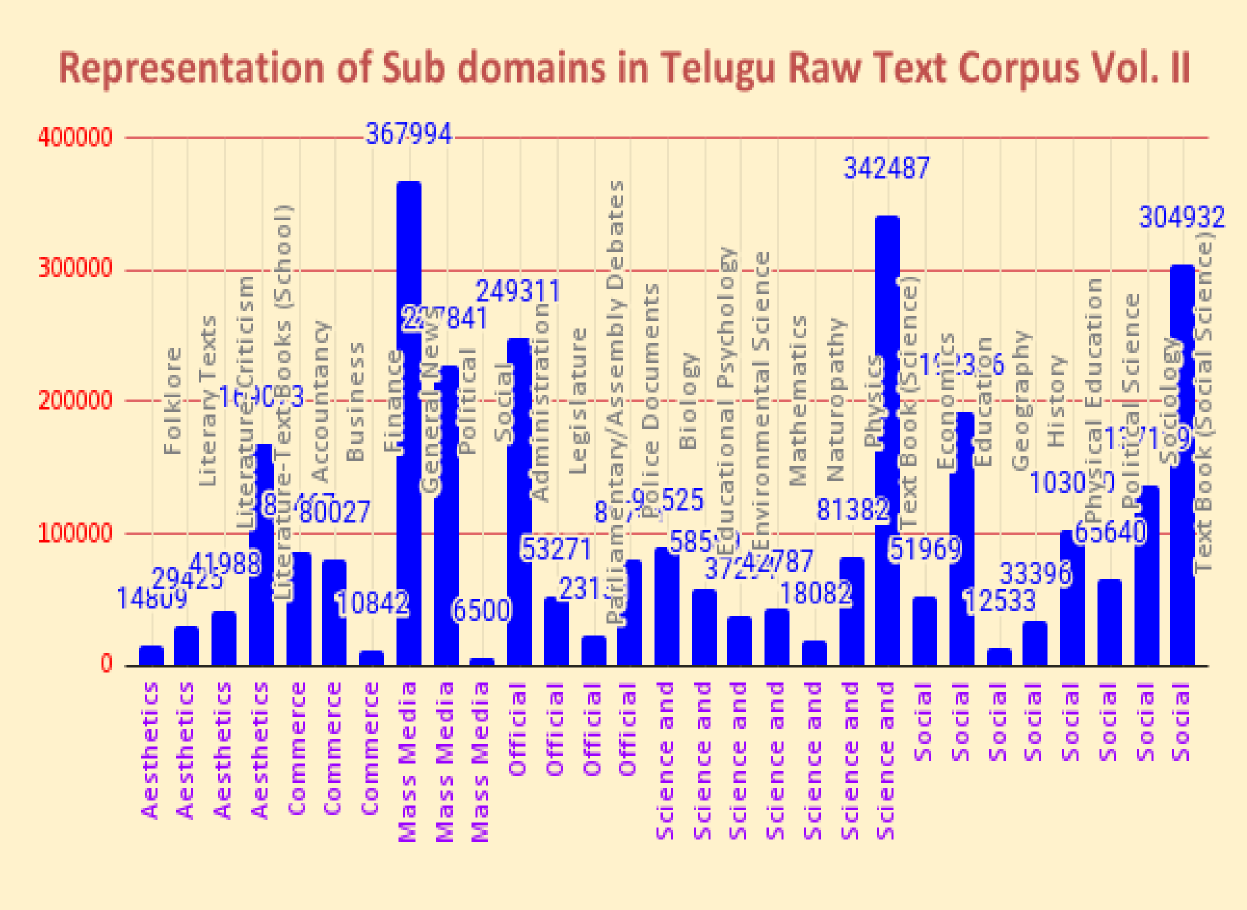

Figure 6: Representation of the subdomains in Telugu Raw Text Corpus Vol. II

6.3 References

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora in Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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.

7 MAITHILI RAW SPEECH CORPUS VOL. II

Shantanu Kumar, Narayan Kumar Choudhary

7.1 INTRODUCTION

The main barrier to language technology development for Indian languages has been a lack of fundamental linguistic resources. One could not even consider speech data when most Indian languages already had text data available. India is one of the foremost multilingual countries where multilingualism is ingrained, and most people speak more than one language, with more than 75 languages having more than one million speakers (as per 2011 Census of India data). As per a study3 by KPMG and Google released in 2017, the internet user base grew at a compound annual growth rate (CAGR) of 41% between 2011 and 2016 to reach 234 million users at the end of 2016, and this trend is likely to continue. It is also estimated that internet users in the Indian language will account for approximately 75% of India's internet user base by 2021.

Even though this is the case, very little technology is available in Indian languages. This is primarily because the institutions responsible for developing new technologies find it either too difficult or not economically feasible to provide linguistic support for a wide range of technology-based applications for Indian languages.

The Indian government has initiated several efforts in response to this problem to supply the core elements that could accelerate the development of language technology in Indian languages. As part of this initiative, the Linguistic Data Consortium for Indian Languages (LDC-IL) was established by the Ministry of Education at the Central Institute of Indian Languages, Mysore.

The goal of LDC-IL was to develop linguistic resources for all Indian languages, with an initial focus on the scheduled languages of India. The language technology development community may deem these resources suitable.

In addition to 22 scheduled languages, LDC-IL has also taken a positive step in its approach towards the mother tongues spoken in India, which is an indication of greater efforts to support and promote linguistic variety in the nation. To acknowledge the significance of the mother tongue, LDC-IL has stepped up its efforts to collect speech data of Maithili. This step towards developing language technology for Indian mother tongues will contribute to the overall enrichment and empowerment of mother tongues and will ensure the continued vitality of the language.

7.2 LDC-IL SPEECH CORPUS

The LDC-IL speech corpus is collected after careful deliberations on what type of speech corpus is required for various types of speech-based linguistic analysis that may suit the multifarious needs of the research and development community.

[3] https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/01/Impact-of-internet-and-digitisation.pdf

After several meetings with the experts from around India and abroad, it was decided that LDC-IL should focus on not just developing a speech corpus for a particular need but rather getting the data that would be useful for various tasks such as ASR, STT, linguistic analysis, speech therapy, and so on.

Keeping this in mind, various types of content were created a priori before the speech recordings took place. The content of these datasets has been prepared in consultation with the experts from the language as well as linguists giving inputs to ensure that no specific sound patterns are missed out.

For example, it has been ensured that the speech datasets contain all the phones and allophones of the language, and ample examples are available in the language to prove their phonemic status in the language. To ensure that the corpus is good for an ASR, the continuous speech is recorded in a natural environment.

7.3 CONTENT TYPE DESCRIPTION

The Maithili raw speech corpus is made up of recordings of native Maithili speakers from various parts of the state of Bihar and Jharkhand, mainly Bhagalpur and neighboring districts, and it represents a wide range of Maithili varieties as they are spoken in various locations by diverse speakers. Though the region is primarily claimed as the Angika-speaking region by several works (Grierson 1907, Sharma 2001) and a few other individual research works, still, based on the government documents and the Census report of 2011 by the Govt. of India, the area is considered as the language variety of Maithili. Hence, the area is referred to as the Maithili-speaking region in this document hereafter.

Below are comprehensive explanations of each of the content types:

7.3.1 **Creative Text (CT)**

The creative text (CT) read speech data includes the recording of a variety of Maithili literary writings. In this content type, the Maithili short stories and essays are read by informants. Any standard descriptive text may be used as creative text. It displays the linguistic preferences of several authors from various regions of the language community, from whence the content was obtained. Some of the stories were taken from textbooks used in public schools.

7.3.2 **Spontaneous Speech (SS)**

The spontaneous speech (SS) data includes recordings of responses to questions from native speakers. The Investigator asked several questions to the informants related to their daily lives to collect the real-time natural speech of a speaker as a spontaneous speech; the answers to these questions had to be given in their own words. LDC-IL made an effort to capture a native speaker's response in a natural speaking style. As any language has a dynamic nature, it tends to change its forms and features within a distance of a few kilometers. To capture the different forms of the language across the geographical regions, the spontaneous speech has been collected as a part of corpus development. These recordings will offer a priceless window into the language's everyday use, capturing the colloquialisms, idioms, and regional vocabularies crucial for creating models for natural language processing.

Three age groups have been chosen for LDC-IL datasets: 16 to 20 years, 21 to 50 years, and above 50 years. An effort has been made to maintain a balanced corpus in terms of age and gender.

7.4 DATASET PREPARATION AND DISTRIBUTION

Each speaker from various age groups recites prompt text extracts of literature, i.e., Creative Text (CT). A minimum of 2000 words are recorded for each speaker in each recording. In terms of data size and time length, read speech makes up the majority of the speech corpora. The dataset also includes Spontaneous Speech (SS) recordings based on conversations and some random topics and questions about government policies, agriculture, state, districts, day-to-day life, culture, historical places, folk tales, and several other things from various fields to maintain the authentic essence of everyday utterances and native terminologies.

A deliberate attempt has been made to maintain distinctiveness in speech corpora by taking a variety of styles into consideration to ensure representativeness in speech corpora. After the regions are identified, speech samples are collected as per the criteria shown in the table below:

Content type	Content size	Content read by each speaker	Age group-wise no. of speaker			Content selection type
			16-20	21-50	50+	
Creative Text	85 Texts	1 Text	07	62	16	Distinct Text
Spontaneous Speech	NA	NA	08	79	34	Prompt questions from various domains and day-to-day life

Table 13: Content-wise Speech Dataset Distribution

7.5 DATA COLLECTION

Fieldwork was conducted in Bihar and Jharkhand, where the Angika variety is spoken. Speech data comprising 122 speakers was collected in these fieldworks. Six investigators namely, Shantanu Kumar, Rupesh Kumar Pandey, Akanksha Tiwari , Jyoti Kumari, Nikhil Kumar, Mukesh Kumar have collected the Data, during November 06 - 13, 2022 from Bhagalpur and neighboring areas of Bihar & Jharkhand.

7.6 Summary of the Corpora

LDC-IL Maithili raw speech corpus vol.II has 206 audio segments with a duration of 109:09:50 (hh:mm:ss).

The data distribution of the speech corpora is shown in the table below:

Content Type	Gender	Female			Male		
	Age Group	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Creative Text-CT	85	06	28	04	01	34	12
Spontaneous Speech-SS	121	07	35	07	01	44	27

Table 14: Audio Segments and their Distribution

Content-wise data distribution and their duration are shown in the table below:

Content Type	**Gender**	**Age Group**	**Duration (hh:mm:ss)**		
Creative Text	Female	16To20	3:19:17	18:59:14	42:46:40
		21To50	13:41:49		
		Above51	1:58:08		
	Male	16To20	0:42:33	23:47:26	
		21To50	17:37:15		
		Above51	5:27:38		
Spontaneous Speech	Female	16To20	3:32:58	26:17:15	66:23:11
		21To50	18:59:32		
		Above51	3:44:45		
	Male	16To20	0:25:02	40:05:57	
		21To50	24:55:06		
		Above51	14:45:48		

Table 15: Content-wise Speech Dataset Distribution and their duration

7.7 References

1. Grierson, GA. 1909, An introduction to the Maithili dialect of the Bihari language as spoken in North Bihar (2 ed.). Calcutta: Asiatic Society of Bengal. pp. xi-xiii.
2. Jha, S. 1958, "The formation of the Maithili language." PhD diss., Luzac London. pp. 5-7
3. Sharma, RM. 2007, Angika Bhasha Ka Dhwanivaigyanik Adhyayan (In Hindi) (II ed.). Shri Tara Geeta Printing Press, Bhagalpur, pp. 130-133

4. https://www.angika.com/p/angika-language.html
5. Choudhary, N. (2019). Linguistic resources for AI/NLP in Indian languages. Mysore: Central Institute of Indian Languages.
6. Ethnologue. (n.d.). hne|ISO-639-3. Retrieved 09 15, 2023, from SIL:https://iso639-3.sil.org/code/hne
7. Office of the Registrar General & Census Commissioner, India (ORGI). (2022, 04 22). Language. Retrieved from Census of India 2011: https://censusindia.gov.in/nada/index.php/catalog/42458
8. Brass, P. R. (2005). Language, Religion, and Politics in North India. iUniverse, Lincoln, NE
9. Choudhary, N. K. (2021). The Indian Repository of Resources for Language Technology. Language Resources & Evaluation (I ed., Vol. 55). Springer. https://doi.org/10.1007/s10579-020-09523-3
10. Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2023. Compendium of LDC-IL Sentence Aligned Speech Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-19411-34-4.

8 DOGRI SPEECH ANNOTATION

Rajesha N., Narayan Kumar Choudhary

8.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Dogri Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Dogri Speech Corpus is available [3] and [4]. LDC-IL Dogri Sentence Aligned Speech Corpus files contain an audio file and its corresponding textual layer. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.

A Typical LDC-IL naming convention for Sentence Aligned Speech data is
'Dogri_Female_16To20_Contemporary_Text-T1_SP-0043_T1-0043-001.wav'

LDC-IL Sentence Aligned Speech corpus for Dogri contains read speech from four content type's viz. contemporary text, creative text, sentences and date format. The contemporary text and creative text are sampled from news and essays/novels respectively. The sentences are a collection of phonetically balanced sentence lists - each speaker has typically recorded 25 sentences randomly selected from this set. Date format contains different way of date formats uttered by the speaker.

8.2 OBSERVATIONS

LDC-IL sentence-level speech annotation strictly follows what the speaker pronounces. The text has been written in the official script of the language and the speech is transcribed as narrowly as the script supports. Even if it is read speech data, there are widespread variations or over corrections when the speaker is reading.

There were also variations in how numbers were pronounced. For example, while reading sports news, the speakers read scores of different sports such as cricket, tennis, etc. in different ways and they deviated from the standardised way of pronouncing the scores. Similarly, there were some errors in reading large numbers such as thousands or lakhs and also in reading decimals, fractions, etc. Most of the speakers faced difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely-used words also influenced the reader's fluency.

8.2.1 PHONETIC ALTERNATION IN DOGRI SPEECH DATA

Read speech has disfluencies like unwanted pauses, elongated syllables, word fragments, self-corrections, and repeated words. Some such disfluency in the recording is given below:

a. Repetition of words

While reading, if the informant observes that the word hasn’t been pronounced in correct or effective manner then normally the speaker repeats a part of that word, the whole word or sometimes even the phrase. Sometimes the speaker also struggles to read the text and keeps repeating when the content seems unfamiliar to him or there may be instances of foreign words or such words which are difficult to pronounce. These are mainly instances of self-correction.

b. False start

False start is a common phenomenon in most of the speakers and for some speakers the frequency increases. Usually, it is the replacement of the first word or a syllable of the word but sometimes speakers start with some other letter as well instead of the actual letter.

E.g.: də-dina

c. Addition and Deletion

An extra vowel or a consonant or a syllable is sometimes added into a word. The sound which already exists in the word might be repeated or a different sound might be inserted into the word.

Deletion or elision of a vowel or a consonant or a syllable from a word is also a common phenomenon attested in the corpus.
E.g.: sət̪i:ʃ > st̪i:ʃ

d. Assimilation and Dissimilation

Speech is a continuous syllabic fragment, so the articulatory organs influence the preceding or following sound. Consonant or vowel is changed to a similar sound because of the influence of a nearby speech segment called assimilation. Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.

e. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.

E.g.: bʱaːrət > paːrt
The original form has been kept in the transcription.

f. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.

E.g.: pət̪ɑ > pɑt̪ɑ

g. Substandard alternation

It has been observed that some speakers have consistently replaced the cluster sounds with nearby or easily articulated sound.

E.g.: kʂet̪ɾ > ket̪ɾ

8.3 SUMMARY OF THE CORPUS

The total duration of Dogri Sentence Aligned Speech Corpus is 08:32:54 (hh:mm:ss) comprising 5,039 audio segments from 61 speakers. The following figures show the distribution of the corpus with respect to gender, age and content type, respectively. The tables that follow gives a break-up of the corpus in terms of recordings obtained from different kinds of texts and also other demographic details as well as the age and gender-wise distribution of all the speakers.

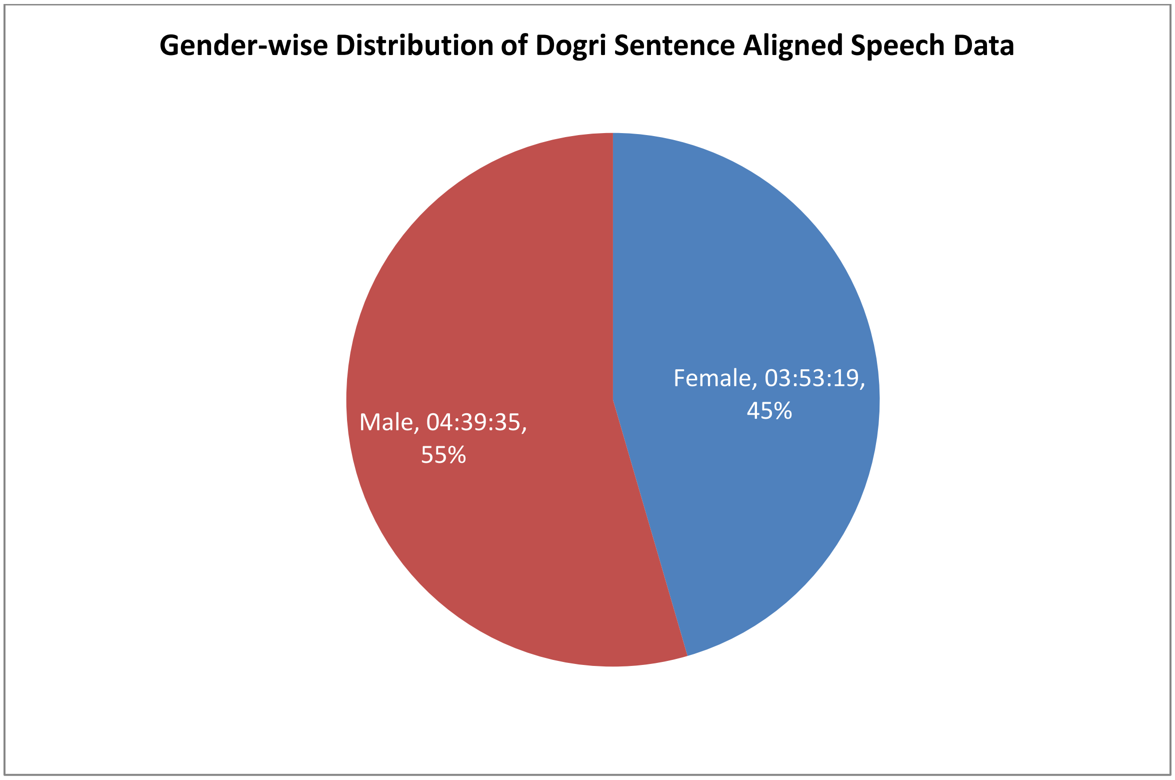

Figure 7: Gender-wise Distribution of Dogri Corpus

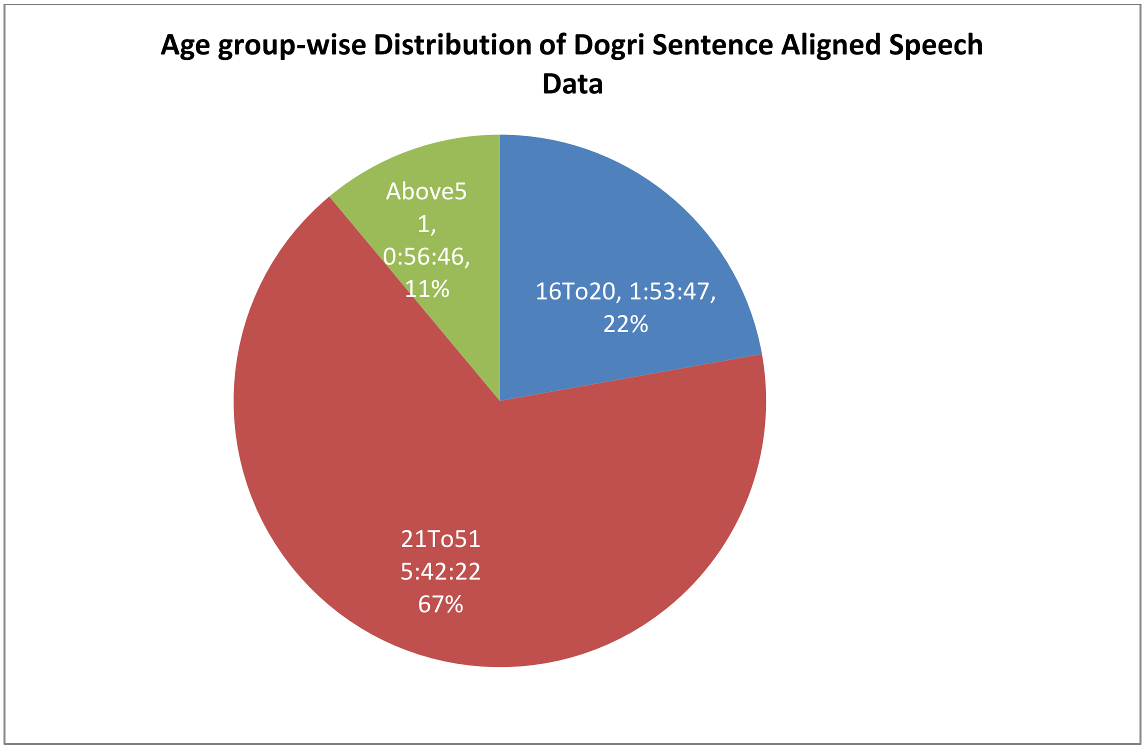

Figure 8: Age-wise Distribution of Dogri Corpus

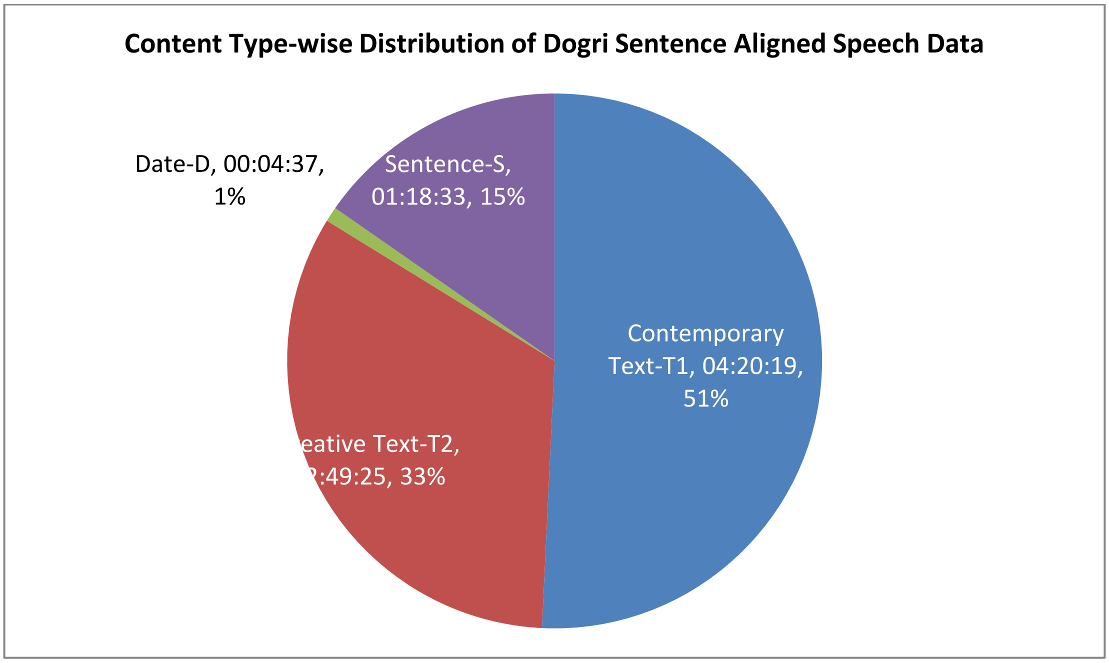

Figure 9: Content Type-wise Distribution of Dogri Corpus

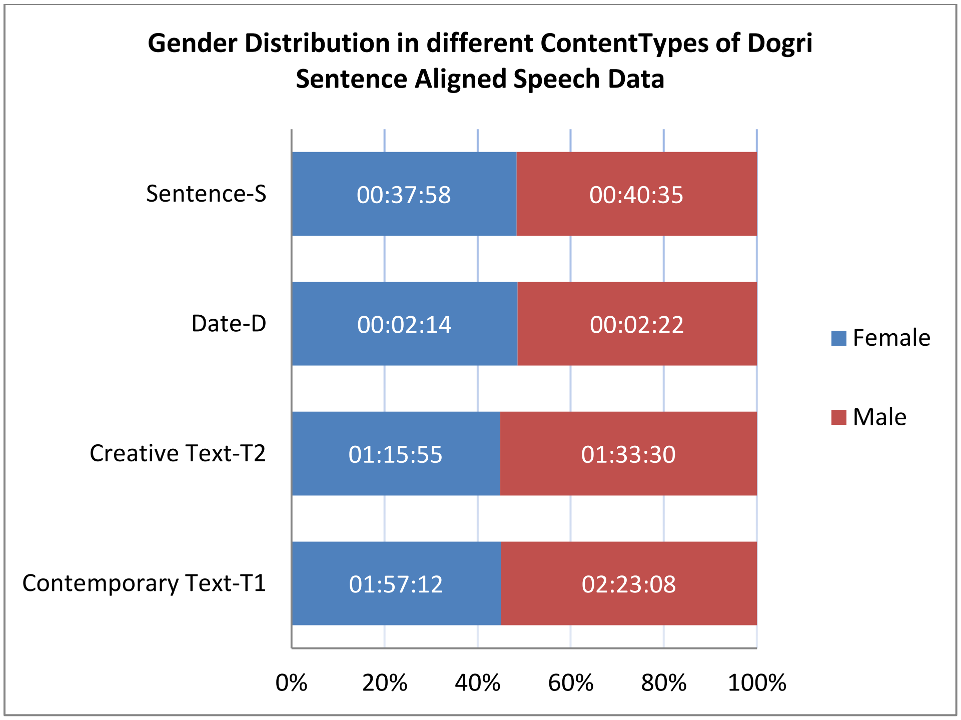

Figure 10: Gender Distribution in different Content Types of Dogri Corpus

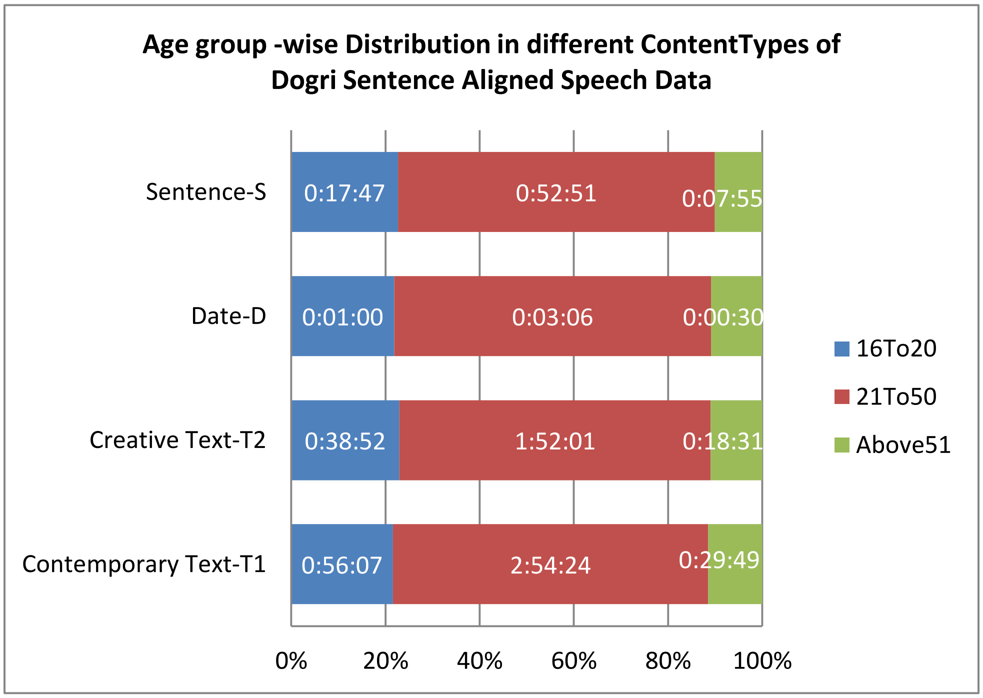

Figure 11: Age group Distribution in different Content Types of Dogri Corpus

8.3.1 DURATION OF DOGRI SENTENCE ALIGNED SPEECH DATA

The table below shows the duration of each of the content types and their distribution across a few factors in Dogri Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	00:27:04.637164	01:57:11.807157	04:20:19.703495
		21To50	01:16:47.081918		
		Above51	00:13:20.088075		
	Male	16To20	00:29:02.080940	02:23:07.896338	
		21To50	01:37:36.474275		
		Above51	00:16:29.341124		
Creative Text-T2	Female	16To20	00:22:08.133904	01:15:54.910421	02:49:24.888109
		21To50	00:46:39.926536		
		Above51	00:07:06.849980		
	Male	16To20	00:16:43.964951	01:33:29.977689	
		21To50	01:05:21.556486		
		Above51	00:11:24.456252		
Date-D	Female	16To20	00:00:35.268053	00:02:14.687596	00:04:37.003685
		21To50	00:01:24.840992		
		Above51	00:00:14.578550		
	Male	16To20	00:00:25.223487	00:02:22.316089	
		21To50	00:01:41.610974		
		Above51	00:00:15.481628		
Sentence-S	Female	16To20	00:10:09.902085	00:37:58.074697	01:18:32.645553
		21To50	00:24:05.503304		
		Above51	00:03:42.669308		
	Male	16To20	00:07:37.388053	00:40:34.570856	
		21To50	00:28:45.039539		
		Above51	00:04:12.143265		

Table 16: Representation of Dogri Sentence Aligned Speech Data Duration

8.3.2 SUMMARY OF SPEAKERS

The table below shows the total number of speakers and their distribution in the Dogri Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	8	5	13
21To50	19	23	42
Above51	3	3	6
Total	30	31	61

Table 17: Distribution of Speakers of Dogri Sentence Aligned Speech Data

8.4 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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 (ed.), Linguistic Resource For AI/NLP in Indian Languages. 2019. Central Institute of Indian Languages, Mysuru. ISBN No 978-81-7343-295-8
4. Choudhary, Narayan (ed.), Compendium of Linguistic Resources in Indian Languages. 2021. Central Institute of Indian Languages, Mysuru. ISBN No.: 978-81-948885-6-7
5. Narayan Kumar Choudhary, Sunil Kumar Choudhary, Rajesha N.,ManasaG., 2021. Dogri Raw Speech Corpus. Central Institute of Indian Languages, Mysore.
6. Ramamoorthy, L., Narayan Choudhary & Sunil Kumar. 2019. A Gold Standard Dogri Raw Text Corpus. Central Institute of Indian Languages, Mysore

9 MAITHILI SPEECH ANNOTATION (TIRHUTA SCRIPT)

Dinesh Mishra, Narayan Kumar Choudhary

9.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Maithili Sentence Aligned Speech Corpus (Tirhuta Script) is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Maithili Speech Corpus is available [3]. LDC-IL Maithili Sentence Aligned Speech Corpus (Tirhuta Script) files contain an audio file and two textual layers in 'Tirhuta' script. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.

A Typical LDC-IL naming convention for Sentence Aligned Speech data is
'Maithili_Female_16To20_Contemporary_Text-T1_SP-0007_T1-0007-001.wav'

LDC-IL Sentence Aligned Speech corpus for Maithili contains read speech from four content type's viz. contemporary text, creative text, sentences and date format. The contemporary text and creative text are sampled from news and essays/novels respectively. The sentences are a collection of phonetically balanced sentence lists - each speaker has typically recorded 25 sentences randomly selected from his set. Date format contains a question uttered by the investigator and the response of the speaker. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalised and the orthographically normalised annotation.

LDC-IL Sentence Aligned Speech corpus for Maithili is available in both Devanagari and Tirhuta scripts. The 'Tirhuta' script was standardised by Unicode in June 2015. Since then there are only a few available digitised Maithili fonts, most of which often representing certain letters/symbols in handwritten style to complete the typeface. As a result, Maithili relies on the Devanagari script. The Tirhuta script is used for producing genealogical records, manuscripts of religious texts, personal correspondence, and printing books. Recognizing the significance of Tirhuta, LDC-IL has annotated and made Maithili speech data available in the script to support its propagation and promotion.

9.2 OBSERVATIONS

LDC-IL sentence-level speech annotation strictly follows what the speaker pronounces to produce the phonetically normalised annotation. The text has been written in the official script of the language and the speech is transcribed as narrowly as the script supports. Even if it is read speech data, there are widespread variations in the pronunciation. For example, speakers from different regions speak the same word in different ways. For example, in Samastipur region few speakers never pronounce /ətʃʰ/ and instead of /ətʃʰ/ they consistently pronounce it as /ətʃʰɪ/.

There were also variations in how numbers were pronounced. For example, while reading sports news, the speakers read scores of different sports such as cricket, tennis, etc. in different ways and they deviated from the standardised way of pronouncing the scores. Similarly, there were some errors in reading large numbers such as thousands or lakhs and also in reading decimals, fractions, etc. Most of the speakers faced difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely-used words also influenced the reader's fluency.

9.2.1 PHONETIC ALTERNATION IN MAITHILI SPEECH DATA

Read speech has disfluencies like unwanted pauses, elongated syllables, word fragments, self-corrections, and repeated words. Some such disfluency in the recording is given below:

h. Repetition of words

While reading, if the informant observes that the word hasn't been pronounced in correct or effective manner then normally the speaker repeats a part of that word, the whole word or sometimes even the phrase. Sometimes the speaker also struggles to read the text and keeps repeating when the content seems unfamiliar to him or there may be instances of foreign words or such words which are difficult to pronounce. These are mainly instances of self-correction.

i. False start

False start is a common phenomenon in most of the speakers and for some speakers the frequency increases. Usually, it is the replacement of the first word or a syllable of the word but sometimes speakers start with some other letter as well instead of the actual letter.

E.g.: b-vɪd̪jaː nɪrd̪ʰaː - nɪrd̪ʰaːrɪt̪

j. Addition and Deletion

An extra vowel or a consonant or a syllable is sometimes added into a word. The sound which already exists in the word might be repeated or a different sound might be inserted into the word.

E.g.: dʒɪmmeːvaːrɪ > dʒɪmmvaːriː raːkʰəl > rəkkʰəl

Deletion or elision of a vowel or a consonant or a syllable from a word is also a common phenomenon attested in the corpus.

E.g.: eːgaːrəɦ > gjaːrəɦ səŋgəʈʰn > səgəʈʰn

k. Assimilation and Dissimilation

Speech is a continuous syllabic fragment, so the articulatory organs influence the preceding or following sound. Consonant or vowel is changed to a similar sound because of the influence of a nearby speech segment called assimilation.

E.g.: tʃəkr > tʃəkk

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: kərɪjaː > kəɳɪjaː

l. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.

E.g.: ɒpəreːʃən > aːpəreːsən ɖaːkʈər > ɖaːgəɖər
The original form has been kept in the transcription.

m. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.

E.g.: unniːs > unnɪs pətʃtʃiːs > pətʃtʃɪs

n. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.

E.g.: ʃaːsən > saːsən ʃɪkʂɪt̪ > ʃitʃtʃʰɪt̪

o. Phone variation

It is the alternative pronunciation of the word, and which does not affect the meaning. Both pronunciations are considered to be in free variations.

E.g.: gaː̃d̪ʱi : > gaːnd̪ʱiː

p. Metathesis

There are instances when the speaker reluctantly speaks with the metathesis alteration. In this case, avoiding the incorrect word, the standard correct word has been transcribed.

E.g. /geːləhaː/is the word pronounced by the speaker whereas the correct form of the word should be /geːlaːh/. So, while transcribing, the correct form has been kept.

9.3 Summary of the Corpus

The total duration of Maithili Sentence Aligned Speech Corpus is 41:54:30 (hh:mm:ss) comprising 21,412 audio segments from 300 speakers. Figures 1, 2 and 3 show the distribution of the corpus with respect to gender, age and content type, respectively. Figure 4 and 5 show gender and age distributions for each content type respectively. Table 1 gives a break-up of the corpus in terms of recordings obtained from different kinds of texts and also other demographic details. Table 2 shows the age and gender-wise distribution of all the speakers.

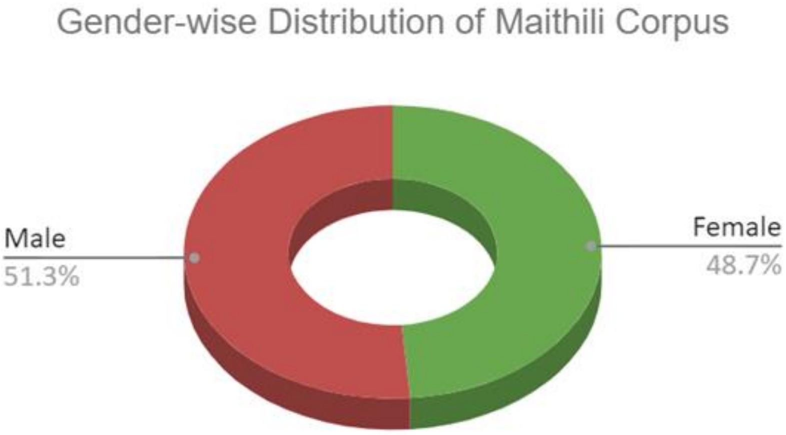

Figure 12: Gender-wise Distribution of Maithili Corpus

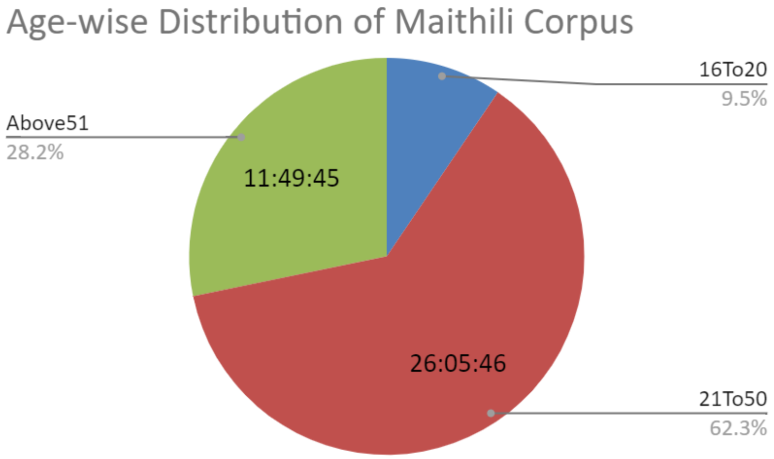

Figure 13: Age-wise Distribution of Maithili Corpus

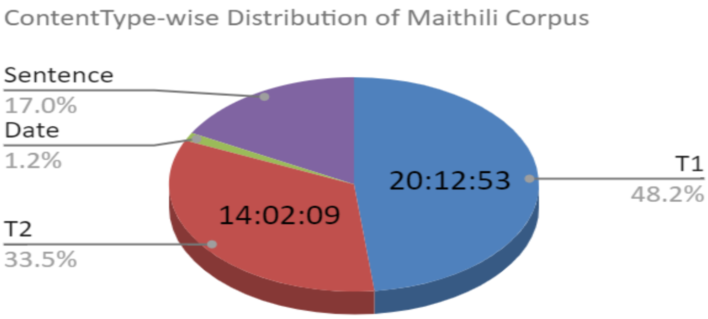

Figure 14: Content Type-wise Distribution of Maithili Corpus

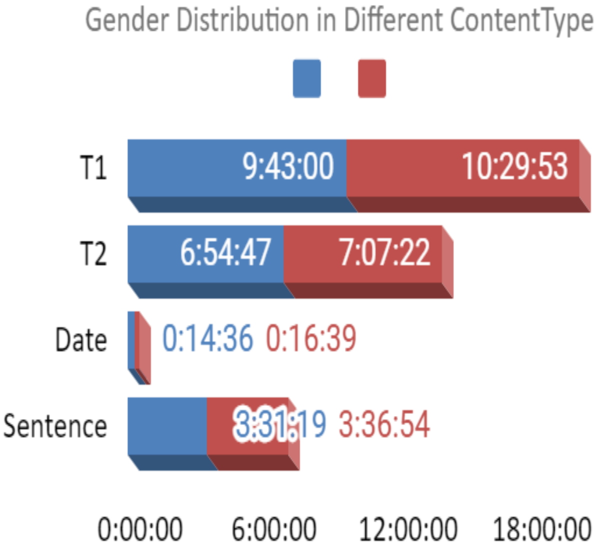

Figure 15: Gender Distribution in different Content Types of Maithili Corpus

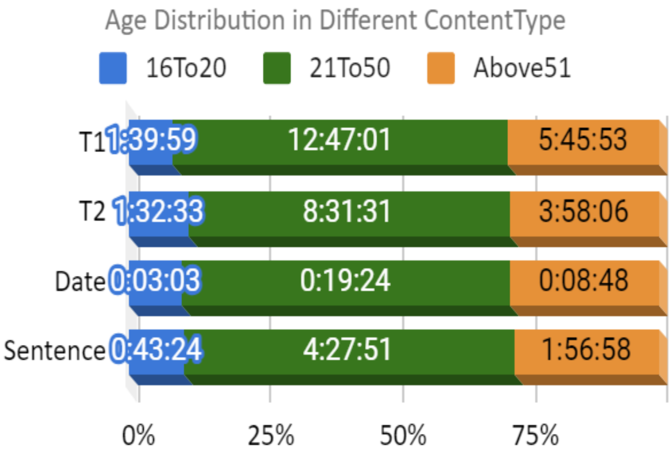

Figure 16: Age Distribution in different Content Types of Maithili Corpus

9.3.1 Duration of Maithili Sentence Aligned Speech Data

The table below shows the duration of each of the content types and their distribution across a few factors in Maithili Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	00:57:07.783085	09:42:59.775607	20:12:52.939966
		21To50	06:07:46.476309		
		Above51	02:38:05.516212		
	Male	16To20	00:42:51.628073	10:29:53.164359	
		21To50	06:39:14.185974		
		Above51	03:07:47.350312		
Creative Text-T2	Female	16To20	00:46:40.027174	06:54:46.948657	14:02:08.984780
		21To50	04:22:01.036929		
		Above51	01:46:05.884554		
	Male	16To20	00:45:52.533406	07:07:22.036123	
		21To50	04:09:29.594624		
		Above51	02:11:59.908092		
Date-D	Female	16To20	00:01:24.594480	00:14:36.302954	00:31:15.019064
		21To50	00:09:32.138534		
		Above51	00:03:39.569940		
	Male	16To20	00:01:38.264070	00:16:38.716110	
		21To50	00:09:51.571019		
		Above51	00:05:08.881021		
Sentence-S	Female	16To20	00:20:30.674311	03:31:19.119820	07:08:13.520495
		21To50	02:16:26.490076		
		Above51	00:54:21.955433		
	Male	16To20	00:22:53.650018	03:36:54.400675	
		21To50	02:11:24.829998		
		Above51	01:02:35.920659		

Table 18: Representation of Maithili Sentence Aligned Speech Data Duration

9.4 Summary of Speakers

The table below shows the total number of speakers and their distribution in the Maithili Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	15	16	31
21To50	93	93	186
Above51	39	44	83
Total	147	153	300

Table 19 : Distribution of Speakers of Maithili Sentence Aligned Speech Data

9.5 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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 (ed.), Linguistic Resource For AI/NLP in Indian Languages. 2019. Central Institute of Indian Languages, Mysuru. ISBN No 978-81-7343-295-8
4. Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh, Dinesh Mishra & Atuleshwar Jha. 2019. Maithili Raw Speech Corpus. Central Institute of Indian Languages, Mysore.
6. Shantanu Kumar, Dinesh Mishra, Rajesha N., Manasa G., Srikanth D., Stephen Fernandes, Nithin S., Narayan Kumar Choudhary, Shailendra Mohan. 2023. Maithili Sentence Aligned Speech Corpus Central Institute of Indian Languages, Mysore. 978-81-19411-96-2.

10 MANIPURI SPEECH ANNOTATION (BENGALI SCRIPT)

Amom Nandaraj Meete., Yumnam Premila Chanu

10.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Manipuri Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL (Ramamoorthy, L. et. Al, 2019). A detailed explanation of the Manipuri Raw Speech Corpus will be available in the Manipuri Speech Data Documentation (Nandaraj, A.M., et. Al, 2019). LDC-IL Manipuri Sentence Aligned Speech files contain an audio file and two textual layers in Manipuri scripts: Bengali and Meetei Mayek. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.

A Typical LDC-IL naming convention for Sentence Aligned Speech data is
'Manipuri_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Manipuri contains read speech from four content type's viz. contemporary text, creative text, sentences and date format. The contemporary text and creative text are sampled from news and essays/novels respectively. The sentences are a collection of phonetically balanced sentence list - each speaker has typically recorded 25 sentences randomly selected from this set. Date format contains a question uttered by the investigator and the response of the speaker. The corpus consists of an audio file for each recording and corresponding textual layers consisting of the phonetically normalised annotation and the orthographically normalised annotation.

10.2 OBSERVATIONS

The LDC-IL sentence-level speech annotation strictly adheres to the speaker's pronunciation to produce a phonetically normalised transcription. The text is written in the language's official script, with the transcription as precise as the script allows. Although the data consists of read speech, pronunciation varies widely. For example, speakers from different regions may pronounce the same word differently. In the Kakching dialect, the word *মানা* /mɑ-nə/ ('by him') exhibits a prominent low tone on the lexical root /mɑ/, which can be transcribed as /mɑ̀-nə/. In contrast, the Sekmai dialect features a long middle tone on the lexical root, transcribed as /mɑ̄-nə/.

There were also variations in how numbers were pronounced. For example, while reading sports news, the speakers read scores of different sports such as cricket, tennis etc. in different ways and they deviated from the standardised way of pronouncing the scores. Similarly, there were some errors in reading large numbers such as thousands or lakhs and also in reading decimals, fractions, etc. Most of the speakers faced difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely-used words also influenced the reader's fluency.

10.2.1 PHONETIC ALTERNATION IN MANIPURI SPEECH DATA

Read speech contains various disfluencies, including unintended pauses, elongated syllables, word fragments, self-corrections, and repetitions. In contrast, Manipuri also exhibits distinct phonological processes, including assimilation, dissimilation, consonant and vowel deletions and additions, as well as consonant strengthening. These differences are exemplified below.

(A) Some Disfluencies

a. Repetition of words

While reading, if the informant observes that the word has been pronounced not in correct or effective manner, then normally the speaker repeats part of the word, whole word or the phrase. Sometimes the speaker was struggling to read the text and repeats when the content is about unfamiliar subjects or there are many foreign words or words which are difficult to pronounce. These are mainly instances of self-correction.

b. False start

False starts are common among most speakers and occur frequently in some. They typically involve the repetition of the first word or syllable, though speakers may occasionally begin with a different letter.

E.g. ম-মতম্ /mə-mətəm/ 'time'; লাঙ্–লাংতক্নবা /lɑŋ- lɑŋtəknəbə/

c. Intended speech

Intended speech occurs when the speaker slows down or fastens up their speech. Typically, it happens at the end of the sentence. It has resulted in inaudible speech in some instances.

For instance, if an audio segment is transcribed as **methəkıdu[də]**, it indicates that **[də]** is not clearly audible. In longer words, certain syllables or phonemes—particularly those in the middle—may be unclear to the listener or inadvertently skipped by the speaker. For example, in **ui[rum]nərəbə**, the middle portion **[rum]** is not distinctly audible.

(B) Phonological Processes

Phonological processes arise as articulatory or perceptual phenomena. When morphemes combine to form words, adjacent segments interact and may undergo changes.In Manipuri, based on the available corpus, these processes can be classified into four categories: assimilation, dissimilation, consonant and vowel deletions and additions, and consonant strengthening.

a. Assimilation

Speech is a continuous stream of syllabic fragments, where articulatory organs influence the preceding sound (progressive assimilation), the following sound (regressive assimilation), or cause two sounds to assimilate to each other (reciprocal assimilation).

Progressive Assimilation

চাপা /cɑpə/ > চাবা/cɑbə/ 'eating'

চাতে /cɑte/ > চাদে/cɑde/ 'not eating'

Regressive Assimilation

পেনলে /penle/ > পেল্লে /pelle/ 'satisfied'

য়েনলে /jenle/ >য়েল্লে /jelle/ 'distributed'

Reciprocal Assimilation

নমকন /nəmkən/ > নঙ্গন /nəŋgən/

শনপন /sənpən/ > শম্বল /səmbən/

b. Dissimilation

Dissimilation is the opposite of assimilation. In Manipuri, certain suffixes beginning with aspirated stops change into unaspirated ones when following roots that end in an unaspirated stop, representing a case of progressive dissimilation.

Examples:

থকথোক /t^{h}əkt^{h}ok/ > থকতোক /t^{h}əktok/

ফুৎখৎ /p^{h}utk^{h}ət/ >ফুৎকৎ /p^{h}utkət/

c. Deletion

The deletion of consonants and vowels simplifies the syllabic structure into a CVCV pattern. As shown below, the lateral /l/ is omitted after a syllable ending in the voiceless velar /k/, while the low vowel /ɑ/ is deleted before the mid vowel /ə/, further streamlining the syllable structure.

Consonant Deletion

কক্লবা /kəkləbə/ >কক’বা /kək’əbə/

হেক্লবা /hekləbə/ > হেক’বা /hek’əbə/

Vowel Deletion

কা অদু /kɑ + ədu/ > কাদু/kɑdu/ ‘that room’

চা অনি /cɑ + ənɪ/ > চনি/cənɪ/ ‘two hundred’

d. Addition

The addition of sounds occurs in both consonants and vowels, though with varying frequency. In Manipuri, consonant addition is rare, whereas vowel addition is more common, particularly in loanwords. Vowels are frequently inserted in initial (prosthesis), medial (epenthesis), and final (epithesis) positions, shaping the phonological structure of borrowed words.

Prothesis

স্কুল /skul/ > ইস্কুল /ɪskul/ ‘school’

ষ্টেন্দর্দ /stendərd/ > ইষ্টেন্দর্দ /ɪstendərd/ ‘standard’

Epenthesis

গ্যাস /gjɑs/ > গীয়াস /gɪjɑs/ ‘gyas’

নর্ক /nərk/ >নরক/nərək/‘hell’

Epithesis

ধর্ম্ম /dhərm/ > ধর্ম্মা /dhərmə/ ‘religion’

তর্ক /tərk/ > তর্কা /tərkə/ ‘argument’

e. Consonant strengthening

Consonant strengthening involves the reinforcement of a segment, often manifesting as the transformation of a non-geminate into a geminate or double consonant. In Manipuri, as observed in the speech corpus, when a suffix beginning with the lateral /l/ is added to roots ending in /p/, /m/, or /ŋ/, it undergoes gemination to strengthen the segment.

Examples:

তপলে /təple/ > তপ্পে /təppe/ ‘have been slow’

খুমলে /k^{h}umle/ > খুম্মে /k^{h}umme/ ‘have cried’

হাংলে /hɑŋle/ > হাংঙে /hɑŋŋe/ ‘have opened’

(C) Colloquial usage

Some speakers have pronounced colloquial forms instead of the standardized forms provided in the prompt sheet. For example:

হায়বসি /hɑɪbəsɪ/ → হায়বসে /hɑɪbəse/ (‘that means’)

মতমসি /mətəmsɪ/ → মতমসে /mətəmse/ (‘this time’ or ‘in this modern time’)

করিনো /kərɪno/ → কৈনো /kəɪno/ (‘what is it’)

The most frequently occurring colloquial functional morphemes in the speech corpus are demonstrative forms, as highlighted below:

Demonstrative Forms

Demonstratives	**Formal**	**Informal**	**Meaning**
Proximal	অসি /əsɪ/	অসে /əse/	‘this’
Distal	অদু /ədu/	অদো /ədo/	‘that’
Pronominal			
Proximal	মসি /məsɪ/	মসে /məse/	‘this’
Distal	মদু /mədu/	মদো /mədu/	‘that’

These variations reflect common colloquial tendencies in spoken Manipuri.

(D) Free Variation in Manipuri Phonology

Free variation occurs when a word has alternative pronunciations without affecting its meaning. In such cases, both variants are phonologically equivalent and interchangeable within a given context.

In Manipuri, this phenomenon is evident in lexical items ending in either /l/ or /r/, where both forms carry the same meaning. This phonetic flexibility, as reflected in the speech corpus, highlights a broader pattern of phonological variability in the language.

E.g.: লৈকোন /ləɪkon/ > লৈকোল /ləɪkol/ ‘garden’; তুরেন /turen/ > তুরেল /turel/ ‘river’

(E) Allophone Realization in Manipuri

Allophones, which are variant forms of a phoneme, occur in complementary distribution. In Manipuri, as observed from the analyzed corpus, four vowel phonemes—/ɪ/, /e/, /u/, and /o/—exhibit allophonic variations.

a. **The Phoneme /ɪ/**

The vowel phoneme /ɪ/ is realized as [jɪ] when it occurs as the initial sound of the second syllable following the vowels /ə/ and /o/:

অইবা /əɪbə/ → অয়িবা [əjɪbə] (‘writer’)

লোইন /loɪn/ → লোয়িন [lojɪn] (‘bodyguard’)

When /ɪ/ follows the vowel /u/, it is realized as [wɪ]:

তৌই /təuɪ/ → তৌৱি [təuwɪ] (‘is done’)

b. **The Phoneme /e/**
The phoneme /e/ is realized as [je] when it appears at the beginning of the second syllable following the vowels /e/ and /u/:
তেএ /tee/ → তেয়ে [teje] (‘is tame’)
পুএ /pue/ → পুয়ে [puje] (‘have carried’)
Additionally, when /e/ follows the diphthongs /əu/ and /ɑu/, it is realized as [we]:
কৌএ /kəue/ → কৌৱে [kəuwe] (‘it is short’)
চাওএ /cɑue/ → চাওৱে [cɑuwe] (‘it is big’)

c. **The Phoneme /u/**
The phoneme /u/ is realized as [wu] when it occurs at the beginning of the second syllable after the central vowel /ə/:
অউবা /əubə/ → অৱুবা [əwubə] (‘one who sees’
When /u/ follows the vowel /ɪ/, it is realized as [ju]:
থিউ /thiu/ → থিয়ু [thju] (‘search’)

d. **The Phoneme /o/**
The phoneme /o/ is realized as [wo] when it appears at the beginning of the second syllable following the central vowel /ə/ and the low vowel /ɑ/:
অওনবা /əonbə/ → অৱোনবা [əwonbə] (‘something that changes’)
When /o/ follows the vowel /ɪ/, it is realized as [jo]:
মীওং /mɪoŋ/ → মীয়োং [mɪjoŋ] (‘type of man’)

This analysis highlights phonetic alternations in Manipuri speech data, following the LDC-IL sentence-level speech annotation system. The phonetic normalization process strictly adheres to the speaker’s pronunciation, ensuring an accurate transcription.

10.3SUMMARY OF THE CORPUS

The **Manipuri Sentence Aligned Speech Corpus**, presented in both **Bengali Script** and **Meetei Mayek**, spans an impressive total duration of 116:34:24 (hh:mm:ss). This extensive dataset is composed of **60,819 meticulously curated audio segments,** contributed by **589 speakers,** showcasing the linguistic richness and diversity of Manipuri speech.

The table below provides a comprehensive breakdown of the duration across different content types, along with its distribution across various key factors within the **Manipuri Sentence Aligned Speech Data,** offering valuable insights into its composition and structure.

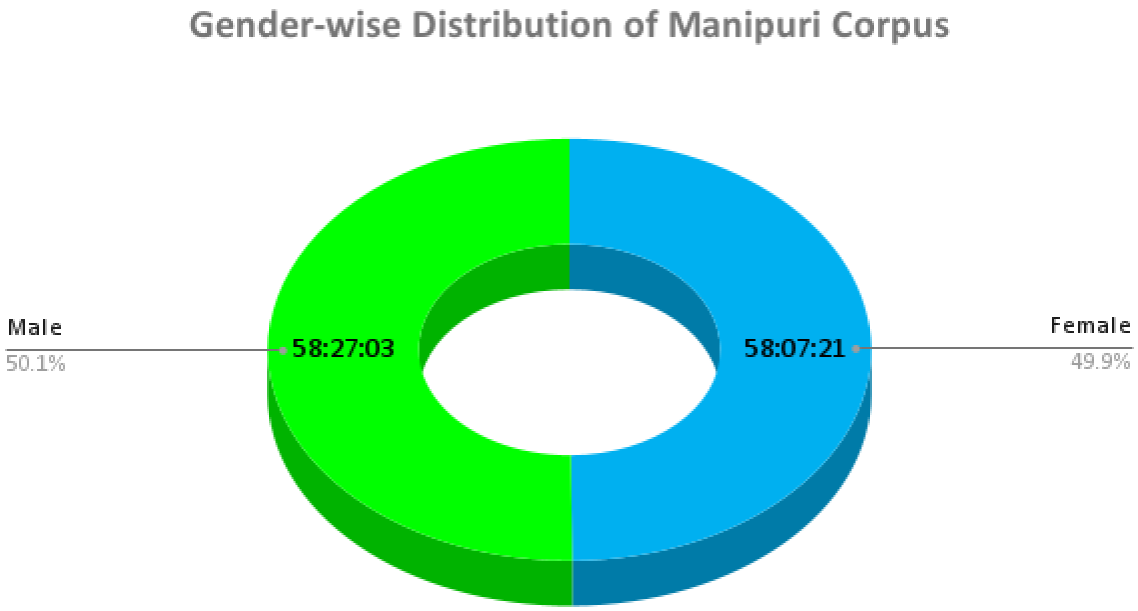

Figure 17: Gender-wise Distribution of Manipuri Corpus

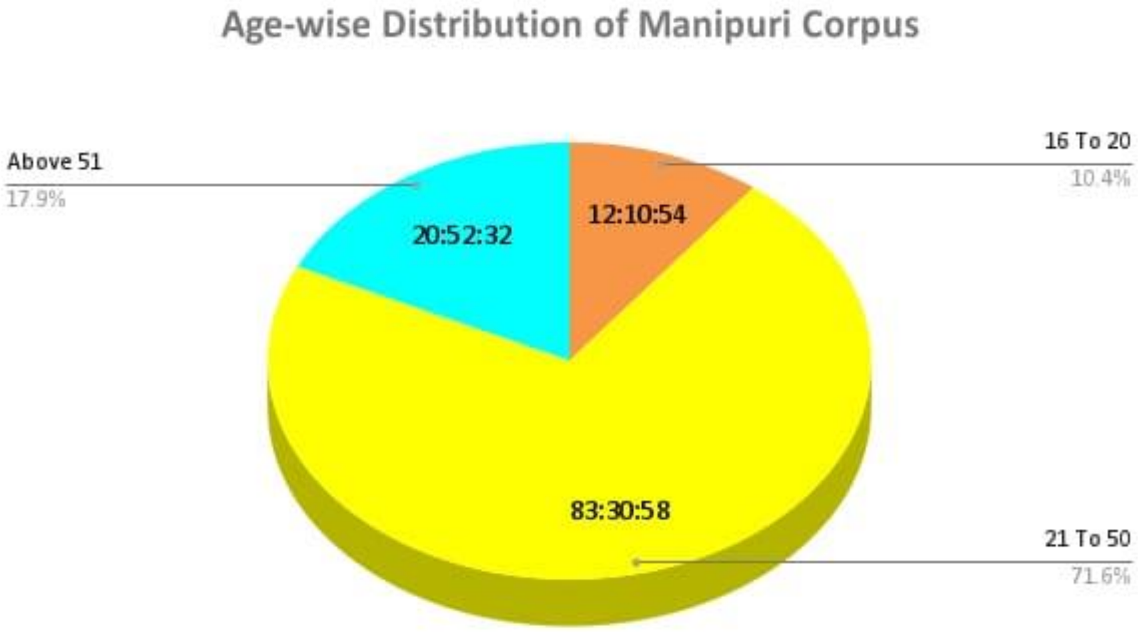

Figure 18: Age-wise Distribution of Manipuri Corpus

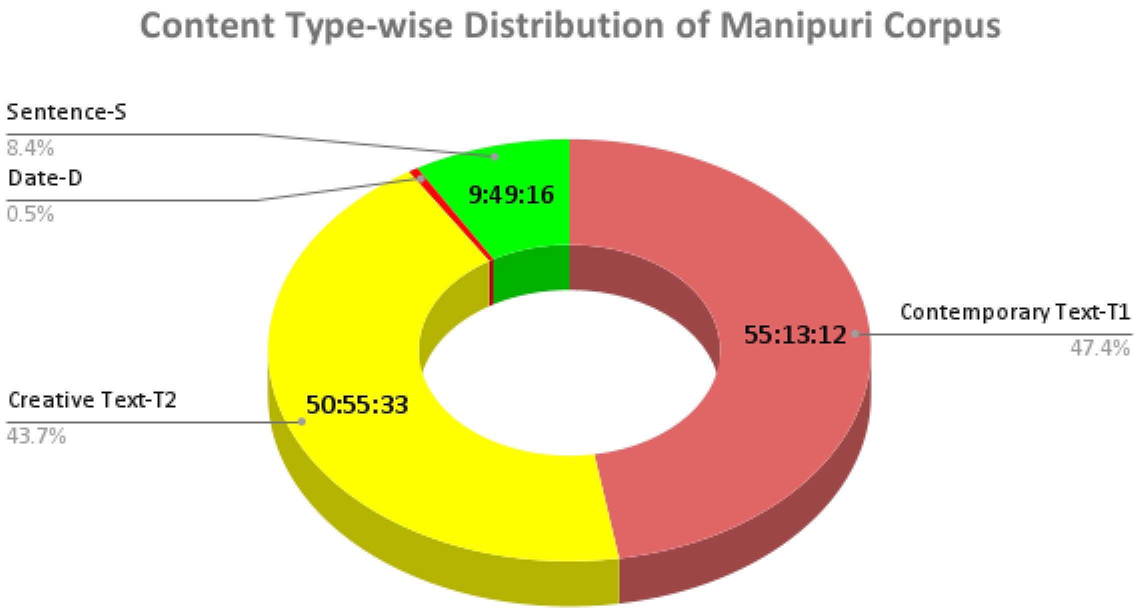

Figure 19: Content Type-wise Distribution of Manipuri Corpus

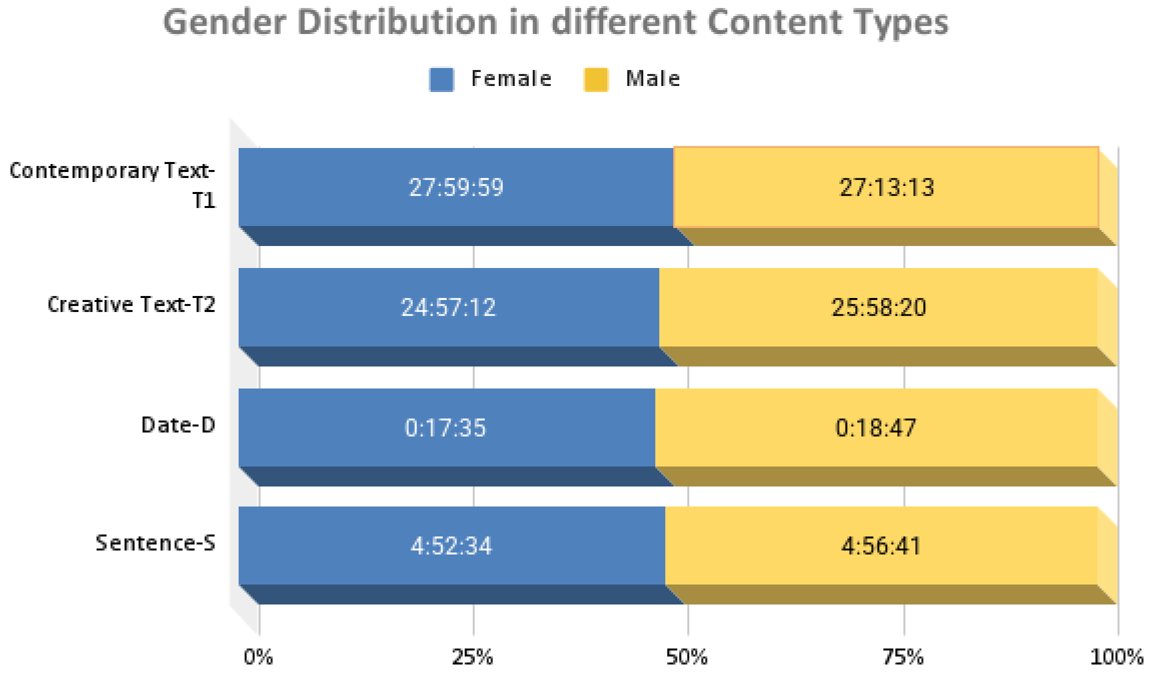

Figure 20: Gender Distribution in different Content Types of Manipuri Corpus

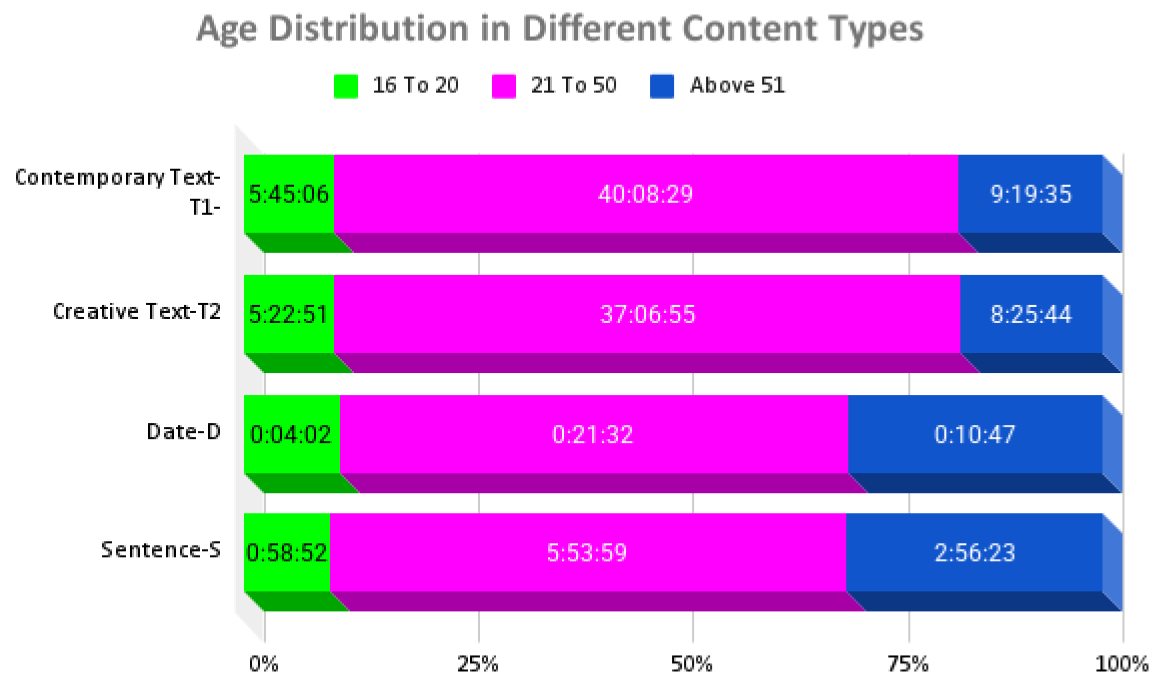

Figure 21: Gender Age Distribution in different Content Types of Manipuri Corpus

10.3.1 Duration of Manipuri Sentence Aligned Speech Data

The table below presents a detailed breakdown of the duration for each content type and its distribution across various factors within the **Manipuri Sentence Aligned Speech Data,** offering valuable insights into its composition.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	02:35:05.004015	27:59:59.297727	55:13:12.425801
		21To50	21:00:47.830039		
		Above51	04:24:06.463674		
	Male	16To20	03:10:01.245593	27:13:13.128074	
		21To50	19:07:42.542715		
		Above51	04:55:29.339766		
Creative Text-T2	Female	16To20	02:05:28.540757	24:57:12.225164	50:55:33.008849
		21To50	18:53:48.854679		
		Above51	03:57:54.829729		
	Male	16To20	03:17:23.378449	25:58:20.783684	
		21To50	18:13:07.432186		
		Above51	04:27:49.973050		
Date-D	Female	16To20	00:01:58.661271	00:17:35.404161	00:36:23.094865
		21To50	00:10:08.222452		
		Above51	00:05:28.520438		
	Male	16To20	00:02:04.265103	00:18:47.690703	
		21To50	00:11:24.120978		
		Above51	00:05:19.304622		
Sentence-S	Female	16To20	00:28:32.915456	04:52:34.476693	09:49:16.126820
		21To50	02:58:55.249092		
		Above51	01:25:06.312145		
		16To20	00:30:19.983125	04:56:41.650127	
	Male	21To50	02:55:04.294377		
		Above51	01:31:17.372624		

Table 20: Representation of Manipuri Sentence Aligned Speech Data Duration

10.4Summary of Speakers

The table below provides a comprehensive overview of the total number of speakers and their distribution within the **Manipuri Sentence Aligned Speech Data.**

Age Group	Female	Male	Total
16To20	29	34	63
21To50	194	187	381
Above51	72	73	145
Total	295	294	589

Table 21: Distribution of Speakers of Manipuri Sentence Aligned Speech Data

10.5References

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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. Amom Nandaraj Meetei, Yumnam Premila Chanu, Rajesha N, Manasa G, Narayan Choudhary & L. Ramamoorthy. 2019. “Documentation of LDC-IL Manipuri Raw Speech Corpus" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 91-103.
5. Ramamoorthy, L., Narayan Choudhary, Amom Nandaraj Meetei, Yumnam Premila Chanu &LongjamAnand Singh. 2019. Manipuri Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

11 Manipuri Speech Annotation (Meetei Mayek)

Amom Nandaraj Meetei, Yumnam Premila Chanu

11.1 Overview of Sentence Aligned Speech Corpus

Manipuri Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL (Ramamoorthy, L. et. Al, 2019). A detailed explanation of the Manipuri Raw Speech Corpus will be available in the Manipuri Speech Data Documentation (Nandaraj, A.M., et. Al, 2019). LDC-IL Manipuri Sentence Aligned Speech files contain an audio file and two textual layers in Manipuri scripts: Bengali and Meetei Mayek. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.

A Typical LDC-IL naming convention for Sentence Aligned Speech data is
'Manipuri_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Manipuri contains read speech from four content types viz. contemporary text, creative text, sentences and date format. The contemporary text and creative text are sampled from news and essays/novels respectively. The sentences are a collection of phonetically balanced sentence list - each speaker has typically recorded 25 sentences randomly selected from this set. Date format contains a question uttered by the investigator and the response of the speaker. The corpus consists of an audio file for each recording and corresponding textual layers consisting of the phonetically normalised annotation and the orthographically normalised annotation.

11.2 Observations

The LDC-IL sentence-level speech annotation strictly adheres to the speaker's pronunciation to produce a phonetically normalised transcription. The text is written in the language's official script, with the transcription as precise as the script allows. Although the data consists of read speech, pronunciation varies widely. For example, speakers from different regions may pronounce the same word differently. In the Kakching dialect, the word □□□ /mɑ-nə/ ('by him') exhibits a prominent low tone on the lexical root /mɑ/, which can be transcribed as /mɑ̀-nə/. In contrast, the Sekmai dialect features a long middle tone on the lexical root, transcribed as /mɑ̄-nə/.

There were also variations in how numbers were pronounced. For example, while reading sports news, the speakers read scores of different sports such as cricket, tennis etc. in different ways and they deviated from the standardised way of pronouncing the scores. Similarly, there were some errors in reading large numbers such as thousands or lakhs and also in reading decimals, fractions, etc. Most of the speakers faced difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely-used words also influenced the reader's fluency.

11.2.1 PHONETIC ALTERNATION IN MANIPURI SPEECH DATA

Read speech contains various disfluencies, including unintended pauses, elongated syllables, word fragments, self-corrections, and repetitions. In contrast, Manipuri also exhibits distinct phonological processes, including assimilation, dissimilation, consonant and vowel deletions and additions, as well as consonant strengthening. These differences are exemplified below.

(E) Some Disfluencies

d. Repetition of words

While reading, if the informant observes that the word has been pronounced not in correct or effective manner, then normally the speaker repeats part of the word, whole word or the phrase. Sometimes the speaker was struggling to read the text and repeats when the content is about unfamiliar subjects or there are many foreign words or words which are difficult to pronounce. These are mainly instances of self-correction.

e. False start

False starts are common among most speakers and occur frequently in some. They typically involve the repetition of the first word or syllable, though speakers may occasionally begin with a different letter.

E.g. □-□□□ /mə-mətəm/ ‘time’; □□□-□□□□□□□ /lɑŋ- lɑŋtəknəbə/

f. Intended speech

Intended speech occurs when the speaker slows down or fastens up their speech. Typically, it happens at the end of the sentence. It has resulted in inaudible speech in some instances.

For instance, if an audio segment is transcribed as **methəkıdu[də]**, it indicates that **[də]** is not clearly audible. In longer words, certain syllables or phonemes—particularly those in the middle—may be unclear to the listener or inadvertently skipped by the speaker. For example, in **ui[rum]nərəbə**, the middle portion **[rum]** is not distinctly audible.

(F) Phonological Processes

Phonological processes arise as articulatory or perceptual phenomena. When morphemes combine to form words, adjacent segments interact and may undergo changes.In Manipuri, based on the available corpus, these processes can be classified into four categories: assimilation, dissimilation, consonant and vowel deletions and additions, and consonant strengthening.

a. Assimilation

Speech is a continuous stream of syllabic fragments, where articulatory organs influence the preceding sound (progressive assimilation), the following sound (regressive assimilation), or cause two sounds to assimilate to each other (reciprocal assimilation).

Progressive Assimilation

□□□ /cɑpə/ > □□□ /cɑbə/ ‘eating’

□□□□ /cɑte/ > □□□□ /cɑde/ ‘not eating’

Regressive Assimilation

□□□□□ /penle/ > □□□□□ /pelle/ ‘satisfied’

□□□□□ /jenle/ > □□□□□ /jelle/ ‘distributed’

Reciprocal Assimilation

□□□□ /nəmkən/ > □□□□ /nəŋgən/ ‘back (body)

□□□□ /sənpən/ > □□□□ /səmbən/

b. Dissimilation ‘fence’

Dissimilation is the opposite of assimilation. In Manipuri, certain suffixes beginning with aspirated stops change into unaspirated ones when following roots that end in an unaspirated stop, representing a case of progressive dissimilation.

Examples:

□□□□□ /t^{h}əkt^{h}ok/ > □□□□□ /t^{h}əktok/ ‘drink off’
□□□□□ /p^{h}utk^{h}ət/ > □□□□□ /p^{h}utkət/ ‘begin to boil’

c. Deletion

The deletion of consonants and vowels simplifies the syllabic structure into a CVCV pattern. As shown below, the lateral /l/ is omitted after a syllable ending in the voiceless velar /k/, while the low vowel /ɑ/ is deleted before the mid vowel /ə/, further streamlining the syllable structure.

Consonant Deletion

□□□□ /kəkləbə/ > □□□□ /kək’əbə/ ‘the cut one’
□□□□□ /hekləbə/ > □□□□□ /hek’əbə/ ‘the plucked one’

Vowel Deletion

□□ □□□ /kɑ + ədu/ > □□□□ /kɑdu/ ‘that room’
□□ □□□ /cɑ + ənɪ/ > □□□ /cənɪ/ ‘two hundred’

d. Addition

The addition of sounds occurs in both consonants and vowels, though with varying frequency. In Manipuri, consonant addition is rare, whereas vowel addition is more common, particularly in loanwords. Vowels are frequently inserted in initial (prosthesis), medial (epenthesis), and final (epithesis) positions, shaping the phonological structure of borrowed words.

Prothesis

□□□□□ /skul/ > □□□□□ /ɪskul/ ‘school’
□□□□□□□□□ /stendərd/ > □□□□□□□□□ /ɪstendərd/ ‘standard’

Epenthesis

□□□□□ /gjɑs/ > □□□□□ /gɪjɑs/ ‘gyas’
□□□□ /nərk/ > □□□ /nərək/‘hell’

Epithesis

□□□□ /dhərm/ > □□□ /dhərmə/ ‘religion’
□□□□ /tərk/ > □□□ /tərkə/ ‘argument’

e. Consonant strengthening

Consonant strengthening involves the reinforcement of a segment, often manifesting as the transformation of a non-geminate into a geminate or double consonant. In Manipuri, as observed in the speech corpus, when a suffix beginning with the lateral /l/ is added to roots ending in /p/, /m/, or /ŋ/, it undergoes gemination to strengthen the segment.

Examples:

□□□□ /təple/ > □□□□ /təppe/ ‘have been slow’
□□□□□ /k^{h}umle/ > □□□□□ /k^{h}umme/ ‘have cried’
□□□□□ /hɑŋle/ > □□□□□ /hɑŋŋe/ ‘have opened’

(G) Colloquial usage

Some speakers have pronounced colloquial forms instead of the standardized forms provided in the prompt sheet. For example:

□□□□□□ /hɑɪbəsɪ/ → □□□□□□ /hɑɪbəse/ (‘that means’)

□□□□□ /mətəmsɪ/ → □□□□□ /mətəmse/ (‘this time’ or ‘in this modern time’)
□□□□□ /kərɪno/ → □□□□ /kəɪno/ (‘what is it’)

The most frequently occurring colloquial functional morphemes in the speech corpus are demonstrative forms, as highlighted below:

Demonstrative Forms

Demonstratives	**Formal**	**Informal**	**Meaning**
Proximal	□□□/əsɪ/	□□□ /əse/	‘this’
Distal	□□□ /ədu/	□□□ /ədo/	‘that’
Pronominal			
Proximal	□□□ /məsɪ/	□□□ /məse/	‘this’
Distal	□□□ /mədu/	□□□/mədu/	‘that’

These variations reflect common colloquial tendencies in spoken Manipuri.

(H) Free Variation in Manipuri Phonology

Free variation occurs when a word has alternative pronunciations without affecting its meaning. In such cases, both variants are phonologically equivalent and interchangeable within a given context.

In Manipuri, this phenomenon is evident in lexical items ending in either /l/ or /r/, where both forms carry the same meaning. This phonetic flexibility, as reflected in the speech corpus, highlights a broader pattern of phonological variability in the language.

E.g.: □□□□□ /ləɪkon/ > □□□□□ /ləɪkol/ ‘garden’; □□□□□ /turen/ > □□□□□ /turel/ ‘river’

(E) Allophone Realization in Manipuri

Allophones, which are variant forms of a phoneme, occur in complementary distribution. In Manipuri, as observed from the analyzed corpus, four vowel phonemes—/ɪ/, /e/, /u/, and /o/—exhibit allophonic variations.

e. **The Phoneme /ɪ/**
The vowel phoneme /ɪ/ is realized as [jɪ] when it occurs as the initial sound of the second syllable following the vowels /ə/ and /o/:
□□□ /əɪbə/ → □□□□ [əjɪbə] (‘writer’)
□□□□ /loɪn/ → □□□□□ [lojɪn] (‘bodyguard’)
When /ɪ/ follows the vowel /u/, it is realized as [wɪ]:
□□□ /təuɪ/ → □□□□ [təuwɪ] (‘is done’)

f. **The Phoneme /e/**
The phoneme /e/ is realized as [je] when it appears at the beginning of the second syllable following the vowels /e/ and /u/:
□□□□ /tee/ → □□□□ [teje] (‘is tame’)
□□□□ /pue/ → □□□□ [puje] (‘have carried’)
Additionally, when /e/ follows the diphthongs /əu/ and /ɑu/, it is realized as [we]:

□□□□ /kəue/ → □□□□ [kəuwe] (‘it is short’)
□□□□□ /cɑue/ → □□□□□ [cɑuwe] (‘it is big’)

g. **The Phoneme /u/**
The phoneme /u/ is realized as [wu] when it occurs at the beginning of the second syllable after the central vowel /ə/:
□□□ /əubə/ → □□□□ [əwubə] (‘one who sees’
When /u/ follows the vowel /ɪ/, it is realized as [ju]:
□□□ /thiu/ → □□□□ [thju] (‘search’)

h. **The Phoneme /o/**
The phoneme /o/ is realized as [wo] when it appears at the beginning of the second syllable following the central vowel /ə/ and the low vowel /ɑ/:
□□□□□ /əonbə/ → □□□□□ [əwonbə] (‘something that changes’)
When /o/ follows the vowel /ɪ/, it is realized as [jo]:
□□□□□ /mɪoŋ/ →□□□□□ [mɪjoŋ] (‘type of man’)

This analysis highlights phonetic alternations in Manipuri speech data, following the LDC-IL sentence-level speech annotation system. The phonetic normalization process strictly adheres to the speaker’s pronunciation, ensuring an accurate transcription.

11.3SUMMARY OF THE CORPUS

The **Manipuri Sentence Aligned Speech Corpus**, presented in both **Bengali Script** and **Meetei Mayek**, spans an impressive total duration of 116:34:24 (hh:mm:ss). This extensive dataset is composed of **60,819 meticulously curated audio segments,** contributed by **589 speakers,** showcasing the linguistic richness and diversity of Manipuri speech.

The table below provides a comprehensive breakdown of the duration across different content types, along with its distribution across various key factors within the **Manipuri Sentence Aligned Speech Data,** offering valuable insights into its composition and structure.

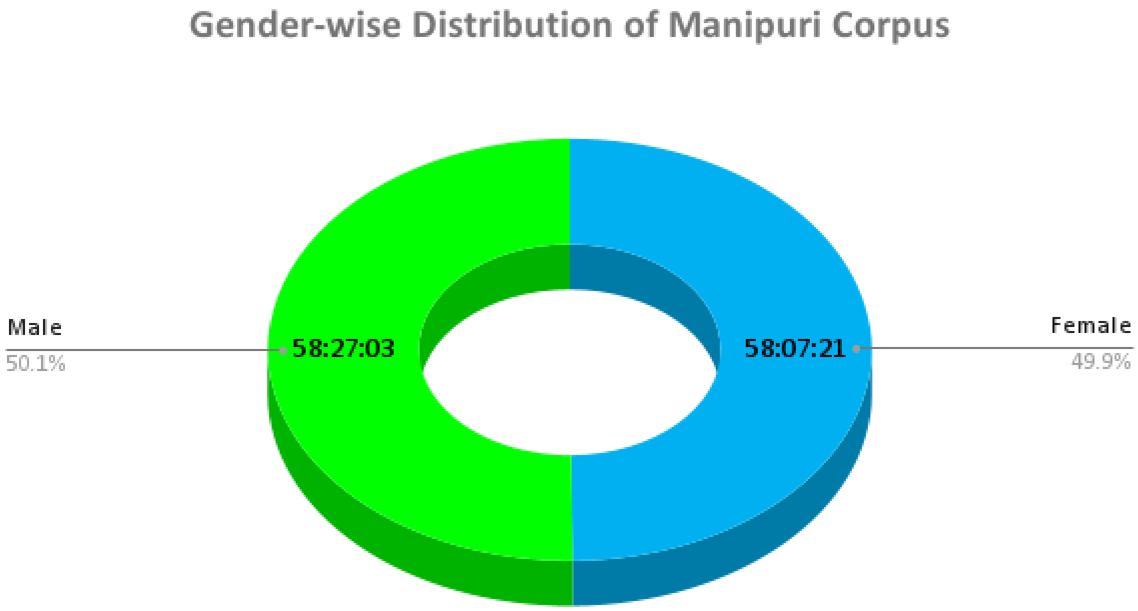

Figure 22: Gender-wise Distribution of Manipuri Corpus

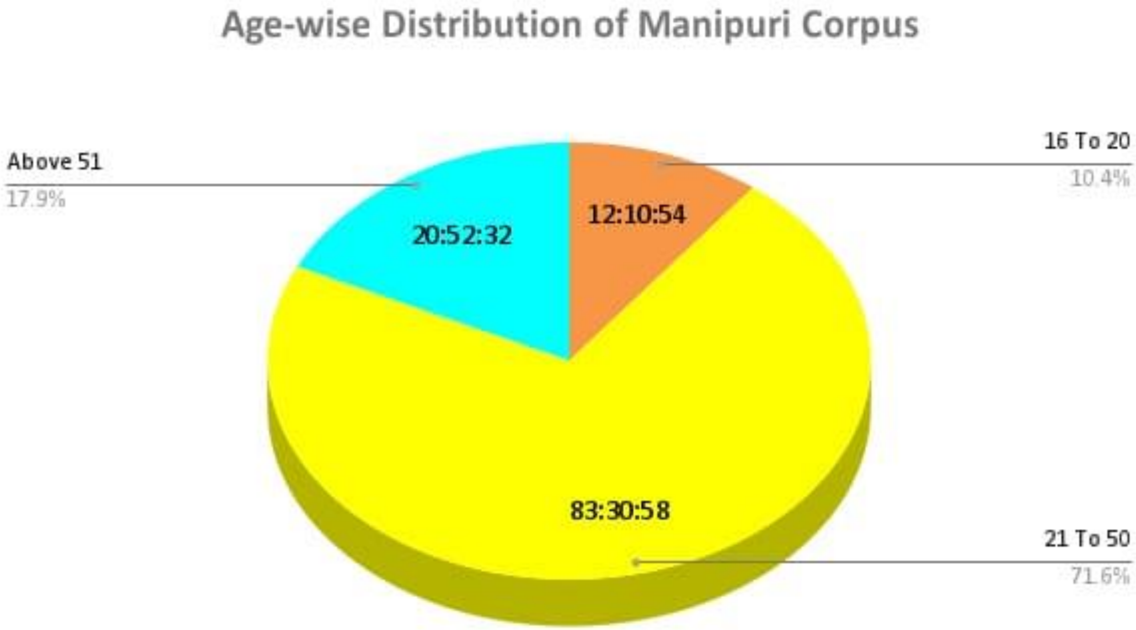

Figure 23: Age-wise Distribution of Manipuri Corpus

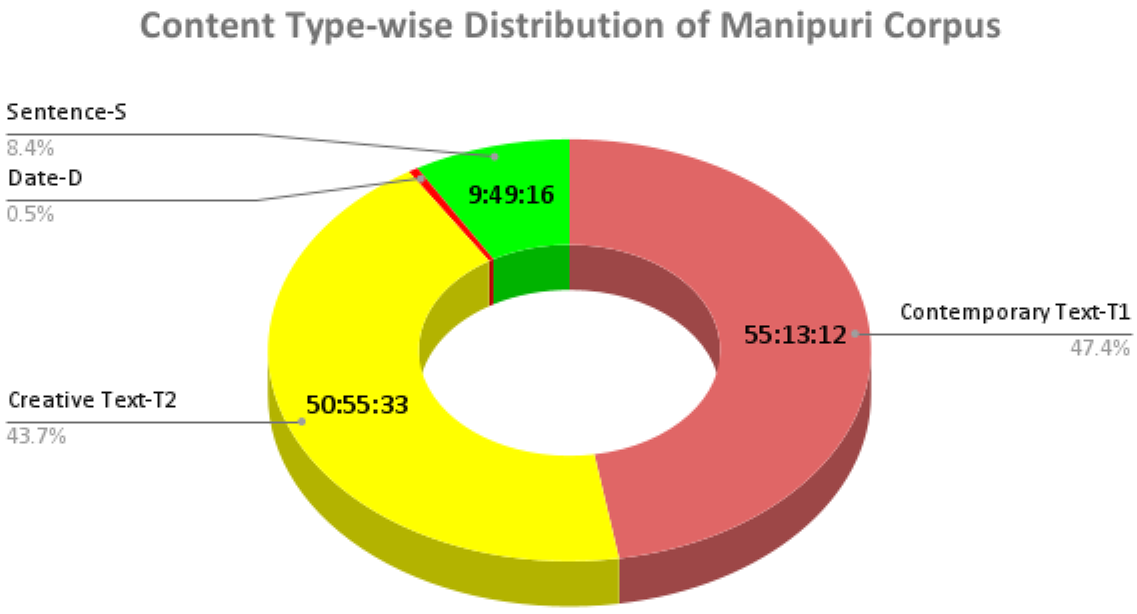

Figure 24: Content Type-wise Distribution of Manipuri Corpus

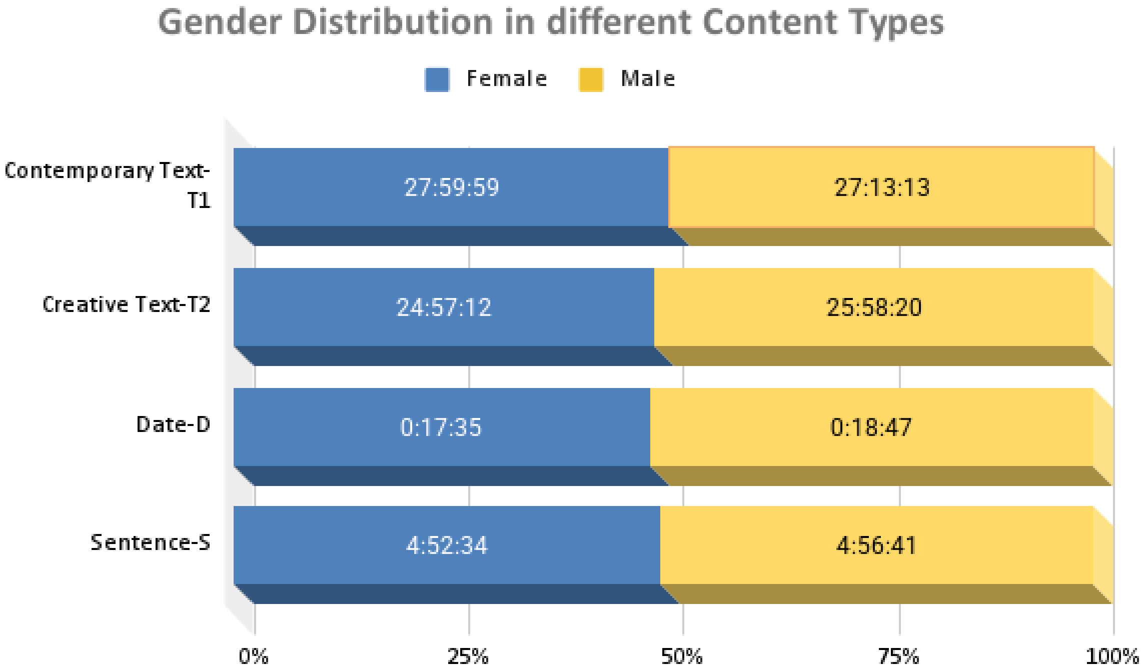

Figure 25: Gender Distribution in different Content Types of Manipuri Corpus

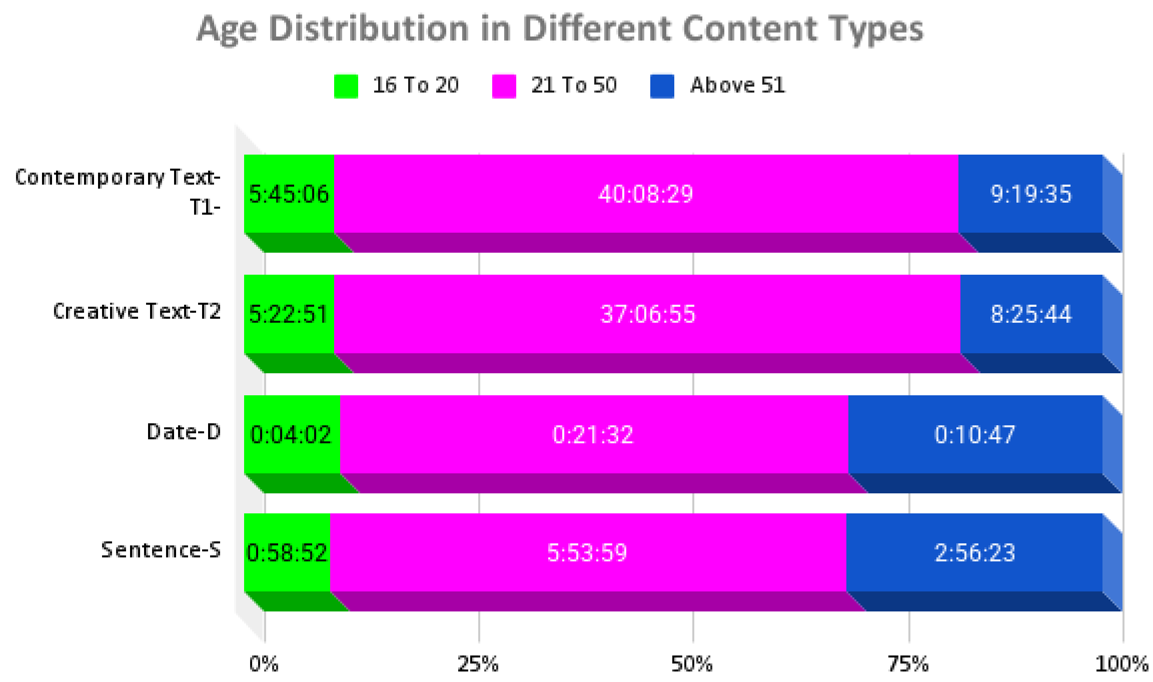

Figure 26: Gender Age Distribution in different Content Types of Manipuri Corpus

11.3.1 DURATION OF MANIPURI SENTENCE ALIGNED SPEECH DATA

The table below presents a detailed breakdown of the duration for each content type and its distribution across various factors within the **Manipuri Sentence Aligned Speech Data,** offering valuable insights into its composition.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	02:35:05.004015	27:59:59.297727	55:13:12.425801
		21To50	21:00:47.830039		
		Above51	04:24:06.463674		
	Male	16To20	03:10:01.245593	27:13:13.128074	
		21To50	19:07:42.542715		
		Above51	04:55:29.339766		
Creative Text-T2	Female	16To20	02:05:28.540757	24:57:12.225164	50:55:33.008849
		21To50	18:53:48.854679		
		Above51	03:57:54.829729		
	Male	16To20	03:17:23.378449	25:58:20.783684	
		21To50	18:13:07.432186		
		Above51	04:27:49.973050		
Date-D	Female	16To20	00:01:58.661271	00:17:35.404161	00:36:23.094865
		21To50	00:10:08.222452		
		Above51	00:05:28.520438		
	Male	16To20	00:02:04.265103	00:18:47.690703	
		21To50	00:11:24.120978		
		Above51	00:05:19.304622		
Sentence-S	Female	16To20	00:28:32.915456	04:52:34.476693	09:49:16.126820
		21To50	02:58:55.249092		
		Above51	01:25:06.312145		
	Male	16To20	00:30:19.983125	04:56:41.650127	
		21To50	02:55:04.294377		
		Above51	01:31:17.372624		

Table 22: Representation of Manipuri Sentence Aligned Speech Data Duration

11.4SUMMARY OF SPEAKERS

The table below provides a comprehensive overview of the total number of speakers and their distribution within the **Manipuri Sentence Aligned Speech Data.**

Age Group	Female	Male	Total
16To20	29	34	63
21To50	194	187	381
Above51	72	73	145
Total	295	294	589

Table 23: Distribution of Speakers of Manipuri Sentence Aligned Speech Data

11.5References

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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. Amom Nandaraj Meetei, Yumnam Premila Chanu, Rajesha N, Manasa G, Narayan Choudhary & L. Ramamoorthy. 2019. “Documentation of LDC-IL Manipuri Raw Speech Corpus" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 91-103.
5. Ramamoorthy, L., Narayan Choudhary, Amom Nandaraj Meetei, Yumnam Premila Chanu &LongjamAnand Singh. 2019. Manipuri Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

12 Punjabi Speech Annotation

Dr. Shalinder Singh, Narayan Kumar Choudhary

12.1 Overview of Sentence Aligned Speech Corpus

Punjabi Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Punjabi Raw Speech Corpus will be available in the Punjabi Raw Speech Corpus (Ramamurthy, L. et. Al, 2019). LDC-IL Punjabi Sentence Aligned Speech files contain an audio file and two textual layers in Gurumukhi script. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.
A Typical LDC-IL naming convention for Sentence Aligned Speech data is shown below.

'Punjabi_Female_16To20_Contemporary Text-T1_SP-0031_T1-0001.wav'

LDC-IL Sentence Aligned Speech corpus for Punjabi contains read speech from four content type's viz. contemporary text, creative text, sentences and date format. The contemporary text and creative text are sampled from news and essays/novels respectively. The sentences are a collection of phonetically balanced sentence lists - each speaker has typically recorded 25 sentences randomly selected from this set. Date format is kept as uttered by the speaker. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalised and the orthographically normalised annotation.

12.2 Observations

LDC-IL sentence level speech annotation strictly follows what the speaker pronounces rather than what is in the prompt sheet. The text has been written in the respective language script and the speech is transcribed as much as the script supports. Two or more different pronunciations can be uttered by the same or different speaker for the same word.

The reading speed differs from reader to reader. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes wrongly uttered large digit numbers like thousands or lakhs, decimal numbers, fractions etc. It is observed that speakers read Cricket score, Tennis score etc. in their own way and very few speakers read it properly. Most of the speakers show difficulty in pronouncing foreign names (other than native language names) which is frequently appearing in sports and political news. Abbreviations and rarely used words interrupt the reader's fluency. All these factors contribute to the complexity in speech which makes it a rather difficult task. Since the dialect of annotator can differ from that of the informant, the annotation process may need repetitive hearing in some cases.

The annotation has to discard the data in particular places where the investigator has communicated with the informant. Some background noise like the sound of a bell, bus horn, other people's conversation, baby crying etc. can be heard in the recording. Since this can be heard along with the voice of the informant, they have to be retained. This slows down the annotation process. Vocal noise of informants like coughing, sneezing etc. can also be observed.

12.2.1 PHONETIC ALTERNATION IN Punjabi SPEECH DATA

Reading speech has in-fluency like unwanted pauses, elongated syllables, word fragments, self-corrections, and repetitive words. When speakers notice what they utter then they suspend their speech and add, delete, or replace words they have already produced. Some fluctuated occurrences were detailed as follows:

a. Repetition of words

While reading, if the informant observes that the word has been pronounced incorrectly or not in an effective manner then the speaker normally repeats fragments of the word or sometimes the whole word or the phrase. Sometimes the speaker struggles to read the text and repeats when the content is a bit unfamiliar or there are many foreign words which are difficult to pronounce.

b. False start

False start is a common phenomenon in most of the speakers and some speakers it is frequent. Usually, it is the repetition of the first word or syllable of the word but sometimes speakers start with some other letter as well.

E.g.: *b - vid ja;* *əmbərs - əmrɪntəsər;* *ɾoɟ - svəɾoɟəgaɾ*

c. Intended speech

Intended speech occurred when the speaker slow down or fasten up their speech. Typically it happened at the end of the sentence or stopping the reading. But it can be observed in the middle of the sentence too. E.g.: *daːktər neɪ ˈmeɪnuː [aːˈraːm]* * *kəˈrən dɪ səlaːh dɪt̪t̪i hɛ*. Here *[aːˈraːm]* * is not properly auduable but native speaker could easily understood the word because of language proficiency. In the long words the middle of the syllable or phone might not be audible to the listener or skip by the speaker.

d. Addition and Deletion

An extra vowel or a consonant or a syllable is added into a word. The letter which is existing in the word or different letter might be added into the word.
E.g.: *sokaː sokkaː* *kɾɪkət- kɾɪkːət.*

Deleting a vowel or a consonant or a syllable from a word is called deletion or elision. It is a common phenomenon when a natural language speaker speaks indistinctly.
E.g.: *məɦɪlə - ˈmeːla;* *vjəvəɦaːrə - beːvaːrə*

e. Assimilation and Dissimilation

Speech is a continuous syllabic fragment, so the articulatory organs influence the preceding or following sound. Consonant or vowel is changed to a similar sound because of the influence of a nearby speech segment called assimilation. E.g. *ʧəkkər - t͡ʃəkkə*

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments. E.g: *tʊhaːɖaː   θoːɖaː*

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardized form written in the prompt sheet.

E.g.: *ɖoːkʈər - ɖaːkʈər,*   *vɪkaːs - bɪkaːs*

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.

E.g.:   Lengthening*:* *məɦaːn - məːɦaːn*
Shortening*:* *əd̪ʰiːjaːt̪mək - əd̪ʰiːjaːt̪mək*

h. Substandard and Shortening

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.

E.g.: *utsaːha > ustaːha;*   *vidʒajoːtsava > vidʒajoːstava;*

i. Phone variation

It is the alternative pronunciation of the word, and which does not affect the meaning. Both pronunciations are considered to be in free variations.

E.g. *kəfən - kəp ən,*   *məzil - mənɟil,*   *məcc əɾ - məcc əɽ*

j. Final vowel modification

In continuous speech the final vowel gets modified at times in some of the speakers:

E.g. *ɾa rəpətl - ɾa rəpət iː*

12.3SUMMARY OF THE CORPUS

Below section is providing the tabular details of the different content types of the Punjabi Sentence Aligned Speech Corpus. These figures may be helpful in tuning the corpus for various purposes of training, testing and evaluating various algorithms as well as provide useful insights into the dataset. The total duration of Punjabi Sentence Aligned Speech Corpus is 52:24:51 (hh:mm:ss) comprising 31,338 audio segments from 449 speakers.

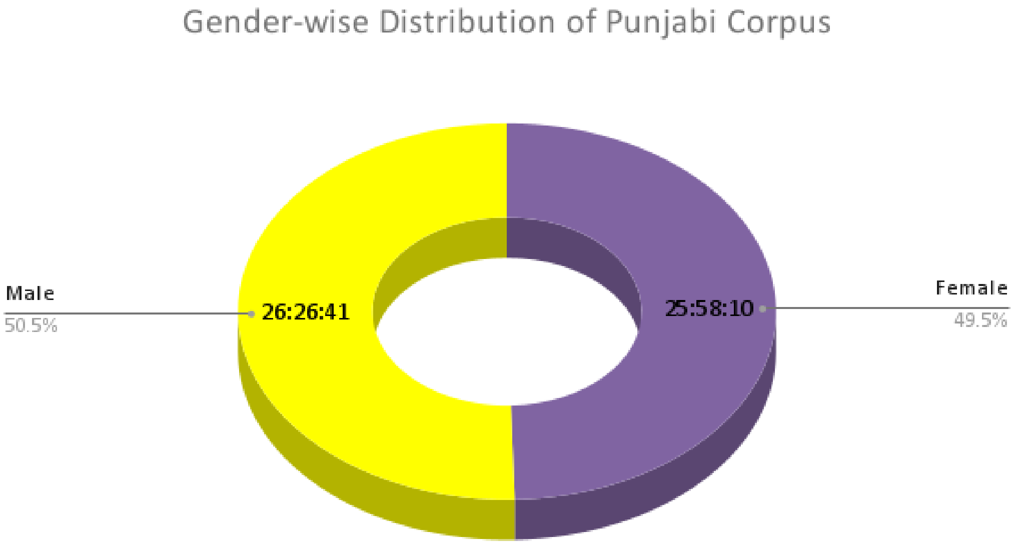

Figure 27: Gender-wise Distribution of Punjabi Corpus

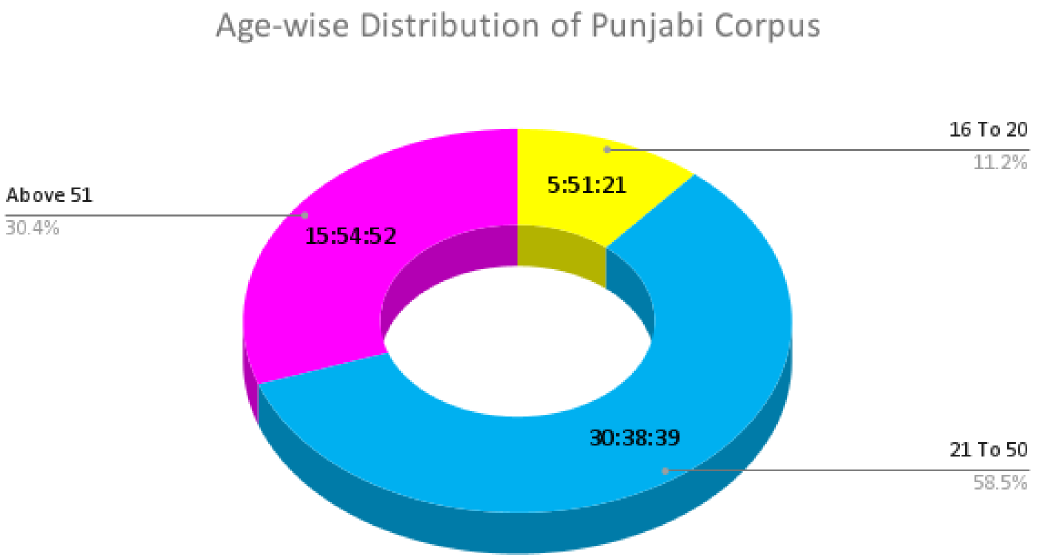

Figure 28: Age-wise Distribution of Punjabi Corpus

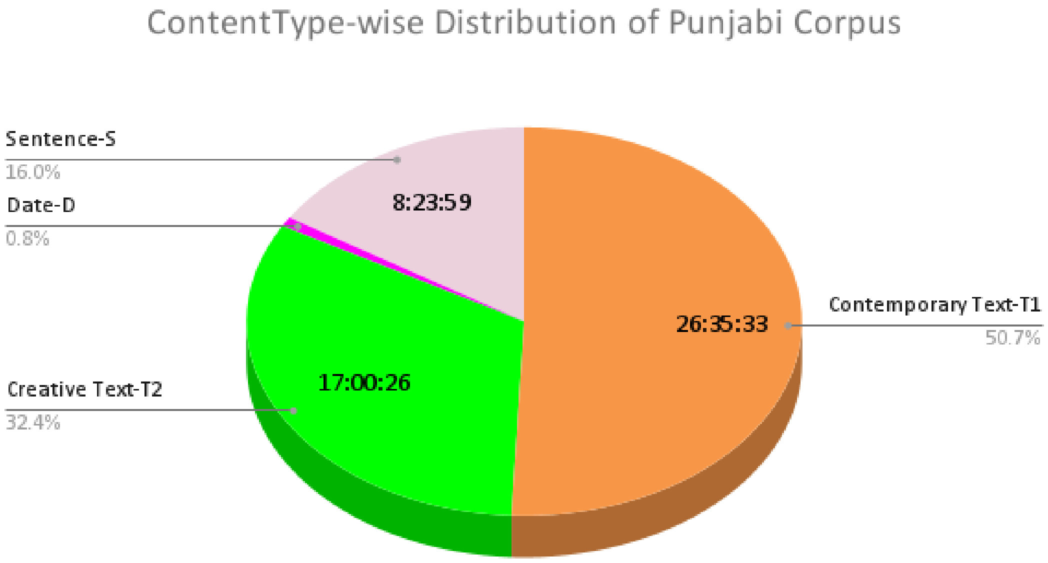

Figure 29: Content Type-wise Distribution of Punjabi Corpus

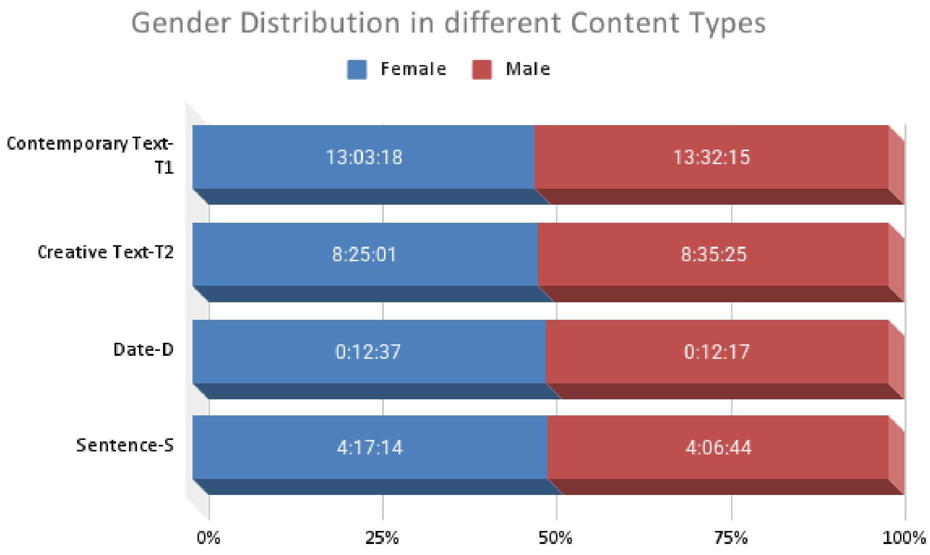

Figure 30: Gender Distribution in different Content Types of Punjabi Corpus

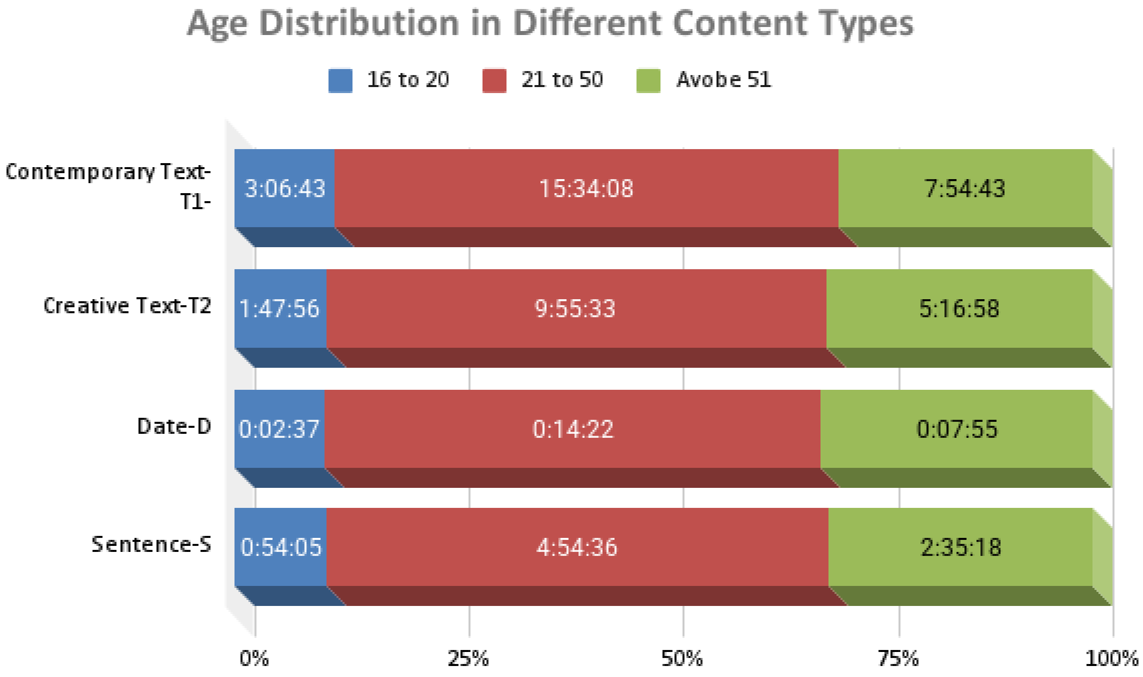

Figure 31: Gender Age Distribution in different Content Types of Punjabi Corpus

12.3.1 DURATION OF PUNJABI SENTENCE ALIGNED SPEECH DATA

The table below shows the duration of each of the content types and their distribution across a few factors in Punjabi Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:44:34.688610	13:03:17.679040	26:35:32.961558
		21To50	07:27:18.357540		
		Above51	03:51:24.632891		
	Male	16To20	01:22:07.587392	13:32:15.282518	
		21To50	08:06:49.817238		
		Above51	04:03:17.877888		
Creative Text-T2	Female	16To20	01:00:41.990339	08:25:00.816673	17:00:26.167396
		21To50	04:54:26.728843		
		Above51	02:29:52.097490		
	Male	16To20	00:47:13.591862	08:35:25.350723	
		21To50	05:01:05.904182		
		Above51	02:47:05.854680		
Date-D	Female	16To20	00:01:26.846645	00:12:36.534899	00:24:53.424891
		21To50	00:07:20.688374		
		Above51	00:03:48.999881		
	Male	16To20	00:01:10.203477	00:12:16.889991	
		21To50	00:07:00.816734		
		Above51	00:04:05.869780		
Sentence-S	Female	16To20	00:29:24.600706	04:17:14.171462	08:23:58.615162
		21To50	02:29:00.253382		
		Above51	01:18:49.317374		
	Male	16To20	00:24:40.033733	04:06:44.443699	
		21To50	02:25:36.220200		
		Above51	01:16:28.189767		

Table 24: Representation of Punjabi Sentence Aligned Speech Data Duration

12.4 SUMMARY OF SPEAKERS

The table below shows the total number of speakers and their distribution in the Punjabi Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	27	23	50
21To50	133	134	267
Above51	65	67	132
Total	225	224	449

Table 25: Distribution of Speakers of Sentence Aligned Speech Data

12.5 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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, Poonam Dhillon & Sarbjeet Kaur. 2019. Punjabi Raw Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Poonam Dhillon, Sarbjeet Kaur & Sandeep Singh. 2019. A Gold Standard Punjabi Raw Text Corpus. Central Institute of Indian Languages, Mysore.

13 Telugu Speech Annotation

Dr. Modugu Kasimbabu, Dr. Narayan Kumar Choudhary

13.1 Overview of Sentence Aligned Speech Corpus

Telugu Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL (Ramamurthy, L. et. Al, 2019). A detailed explanation of the Telugu Raw Speech Corpus will be available in the Telugu Raw Speech Corpus. (Kavitha Lenin. et. Al, 2018). LDC-IL Telugu Sentence Aligned Speech files contain an audio file and two textual layers in Telugu script. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.
A Typical LDC-IL naming convention for Sentence Aligned Speech data is ‘Telugu _Female_16To20_Contemporary Text-T1_SP-0031_T1-0084-001.wav’

The speech is annotated on the basis of specific language syllable structure. The words are labeled manually to the corresponding wave. LDC-IL Sentence Aligned Speech corpus contains four content types such as contemporary text, creative text, sentences and date format. The contemporary text and creative text are recordings of news and essays/novels. Each speaker has uttered typically 25 sentences randomly selected from phonetically balanced sentences list of LDC-IL speech data set. Date format content type contains date format uttered by the speaker. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalized and the orthographically normalized annotation.

13.2 Observations

LDC-IL sentence level speech annotation strictly follows what the speaker pronounces rather than what is in the prompt sheet. The text has been written in the respective language script and the speech is transcribed as much as the script supports. Two or more different pronunciations can be uttered by the same or different speaker for the same word. Even if it is read from speech data, the dialect variation drastically influences the pronunciation. Therefore, speakers from different regions speak the same word in different ways. For e.g. In Costal Andhra region speakers tend to pronounce the years in the connotation of hundreds whereas the other regions prefer in thousands. For example ‘*pad̪ɪheːnu vən̪d̪alu, pəd̪d̪en̪ɪmɪd̪ɪ vən̪d̪alu*’ The Costal Andhra Regions it will be pronounced as ‘*vejjɪ ajɪd̪uvən̪d̪alu, vejjɪ enɪmɪd̪ɪvən̪d̪alu’*. Another noticeable thing is Rayalaseema region people will pronounce the 15 like *‘pad̪ahaid̪u – pad̪aɪd̪u’*. The same number the Costal Andhra and Telangana region people pronounce *‘pad̪ɪheːnu’*.
The reading speed differs from reader to reader. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes wrongly uttered large digit numbers like thousands or lakhs, decimal numbers, fractions etc. It is observed that speakers read Cricket score, Tennis score etc. in their own way and very few speakers read it properly. Most of the speakers show difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely used words interrupt the reader’s fluency. All these factors contribute to the complexity in speech which makes it a rather difficult task. Since the dialect of the annotator can differ from

that of the informant, the annotation process may need repetitive hearing in some cases. The annotation has to discard the data in particular places where the investigator has communicated with the informant. Some background noise like the sound of a bell, bus horn can be heard in the recording. Since this can be heard along with the voice of the informant, they have to be retained. This slows down the annotation process. Vocal noise of informants like coughing, sneezing etc. can also be observed.

13.2.1 PHONETIC ALTERNATION IN TELUGU SPEECH DATA

Reading speech has in-fluency like unwanted pauses, elongated syllables, word fragments, self-corrections, and repetitive words. When speakers notice what they utter then they suspend their speech and add, delete, or replace words they have already produced. Some fluctuated occurrences were detailed as below:

a. Repetition of words

While reading, if the informant observes that the word has been pronounced not in correct or effective manner then normally the speaker repeats part of the word, whole word or the phrase. Sometimes the speaker was struggling to read the text and repeats when the content is about unfamiliar subjects or there are many foreign words or words which are difficult to pronounce. These are mainly instances of self-correction.
For example: *svaːt̪an̪t̪rja d̪ɪnoːt̪savəm* - *svasan̪t̪ra d̪ɪnoːccəva; d͡ʒeːms əɳɖərsən* - *d͡ʒɪmmiː aːɖərsən*

b. False start

False start is a common phenomenon in most of the speakers and some speakers it is frequent. Usually it is the replacement of the first word or syllable of the word but sometimes speakers start with some other letter as well.
E.g.: *pɛccərɪllina* - *hɛccərɪllina; eʈocciː-heccoːʈəciː;* *ɪŋkaː* - *ɪka; keːʈaːɟɪmpu* - *keʈaːɟɪmpu*

c. Intended speech

Intended speech occurs when the speaker slows down or fastens up their speech. Typically, it happens at the end of the sentence. It has resulted in inaudible speech in some instances. If the text is annotated as '*iː mədʰja cɪnnəcɪnnə pəʈʈəɳaːləloː heccərɪllut̪unnə marokə amʃəm saːhit̪iːsəm[st̪ʰəla]* mad̪ʰjə ragɪleː kəkʂalu d̪veːʂaːlu*' shows *[st̪ʰəla]** is not properly audible but native speakers could easily understand the word because of language proficiency. In the long words the middle of the syllable or phone might not be audible to the listener or skip by the speaker. i.e. In, '*rat̪ʰəmu[nu]* d̪eːvunɪ uːreːgicəɖaːnɪki əlaːgeː puːrvəkaːləmloː [o]*kə vaːhənəmgaː raːɟulu vaːɖeːvaːrəni t̪elust̪un̪d̪ɪ*' the middle pair is not audible.

d. Addition and Deletion

An extra vowel or a consonant or a syllable is added into a word. The letter which is existing in the word or different letter might be added into the word.

E.g.: *aːd̪ʰɪpət̪jəvaːd̪ulu* - *aːd̪ʰɪpət̪jaːd̪ɪvaːd̪ulu; kəʈʰɪnət̪vaːnnɪ* - *kəʈʰiːnət̪vaːnnɪ ; aːcaːrəʋjəʋahaːraːlu* > *aːcaːrjəʋjəʋahaːraːlu*

Deletion or elision of a vowel or a consonant or a syllable from a word is also a common phenomenon attested in the corpus.

E.g.: *ɪŋkaː* > *ɪkə ;* *keːʈaːjɪmpu* > *keʈaːjɪmpu ;* *laːnsɛʈ* > *laːnɛʈ* *sprʊhə* > *srʊhə* *rasaːsvaːdanalaku* > *rasvaːd̪alaku*

e. Assimilation and Dissimilation

Speed is a continuous syllabic fragment, so the articulatory organs influence the preceding or following sound. Consonant or vowel is changed to a similar sound because of the influence of a nearby speech segment called assimilation.
E.g.: *muːɖunaːɭləmuccəʈagaː* > *muːnnaːɭləmuccəʈagaː ; bad̪an̪aːmu* > *bəd̪n̪aːmu*
palukut̪uː > *palukut̪uː*

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: *pəd̪d̪ɛmd̪ɪ* > *pəɟɟemd̪ɪ ;* *pəd̪n̪aːlugu* > *pəd̪n̪aːlgu, pəɖnaːlgu ;*

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.: *keːʈaːjɪmpu* > *keʈaːjɪmpu ; ennoːrɛʈlə* > *ennoːreːʈlə ;*

samaraːniki > *samaːraːniki; əʋɪʃʋaːsəm* > *aːʋɪʃʋaːsəm*

h. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.
E.g.: *prauɖʰaməɪna* > *prauɖaməɪna; boːd̪ʰanəloː* > *boːd̪anəloː ; svəccʰən̪d̪əŋgaː* > *svəccən̪d̪əŋgaː grən̪t̪ʰaːlajəmloː* > *grən̪d̪aːlajəmlo*

i. Interchange of Voiced fricative with Vowels

It is observed that some informants interchange the Voiced fricatives (h) with vowels, this is more observed in Rayalaseema and Telangana region.
E.g.: Vowel in place of Voiced fricative.
heccərɪkət̪oː > *eccərɪkət̪oː ; halən̪t̪əŋgaː* > *ələn̪t̪əŋgaː ; həkkunɪ* > *əkkunɪ ; hərɪmcaːru* > *ərɪmcaːru ; hukuːm* > *ukuːm*

Eg. Voiced fricative in place of Vowel

əmməjjaː > həmməjjaː ;   *əkkəraleːd̪u > həkkəraleːd̪u*

j. Interchange of Voiced and voiceless

Telugu has voiced and voiceless consonants; some speakers have pronounced voiced consonants as voiceless or vice versa in some instances.
E.g.: Voiceless in place of Voiced Consonant: *kəljaːɳaːnɪki > kəɭjaːɳaːnɪki ; kʰaːkiːnɪ > kʰaːkʰiːnɪ*

E.g.: Voiced in place of Voiceless Consonant: *neːpət̪ʰjəmloː> neːpəd̪jəmloː;*
kʰəɖgaːnɪki > kəɖgaːnɪki;

k. Interchange of Aspirated to unaspirated

Speakers tend to pronounce aspirated letters in unaspirated, and vice versa across all dialects. Aspirated to unaspirated is more commonly observed in the Rayalaseema region speech.
E.g.: Aspirated in place of unaspirated Consonant:

bʰəɟənaloː > bəɟənaloː ;   *ulləŋgʰɪmcɪnən̪d̪uku > ulləŋgɪmcɪnən̪d̪uku;*
bʰuːmɪki > buːmɪki ;   *kət̪ʰaloː > kət̪aloː*   *viːd̪ʰɪloː > viːd̪ɪlo*;

l. Interchange of Voiceless fricatives

Telugu has three voiceless fricatives namely, శ [ɕ] which is voiceless alveolo-palatal fricative, voiceless retroflex fricative ష [ʂ] and voiceless dental fricative స [s]. It is observed that some informants interchange the Voiceless fricatives.

ʃrəd̪d̪ʰaːsəkt̪ulat̪oː > srəd̪d̪aːʂ̪əkt̪ulat̪oː ; ceːsaːru > ceːʃaːru ; ʂəɳmukʰaːlu > səɳmukaːlu ; ʃməʃaːnəm > sməʃaːnəm ; nɪʂeːd̪ʰaːnnɪ > nɪseːd̪aːnnɪ ; ʃveːt̪aku > sveːt̪aku ; pəuruʂaːnɪki > pəurusaːnɪki ; pəʃuvu > pəsuvu ; ʃud̪d̪ʰəŋgaː > sud̪d̪ʰəŋgaː ; nɪʂpaːkʂɪkəŋgaː > nɪspaːkʂɪkəŋgaː ; nɪrd̪eːʃɪmcɪna > nɪrd̪eːsɪmcɪnə ; pʰaːsɪɟaːnɪki > pʰaːʃɪɟaːnɪki;

13.3 SUMMARY OF THE CORPUS

The total duration of Telugu Sentence Aligned Speech Corpus is 15:38:53 (hh:mm:ss) comprising 9,548 audio segments from 80 speakers. The table below shows the duration of each of the content types and their distribution across a few factors in Telugu Sentence Aligned Speech Data.

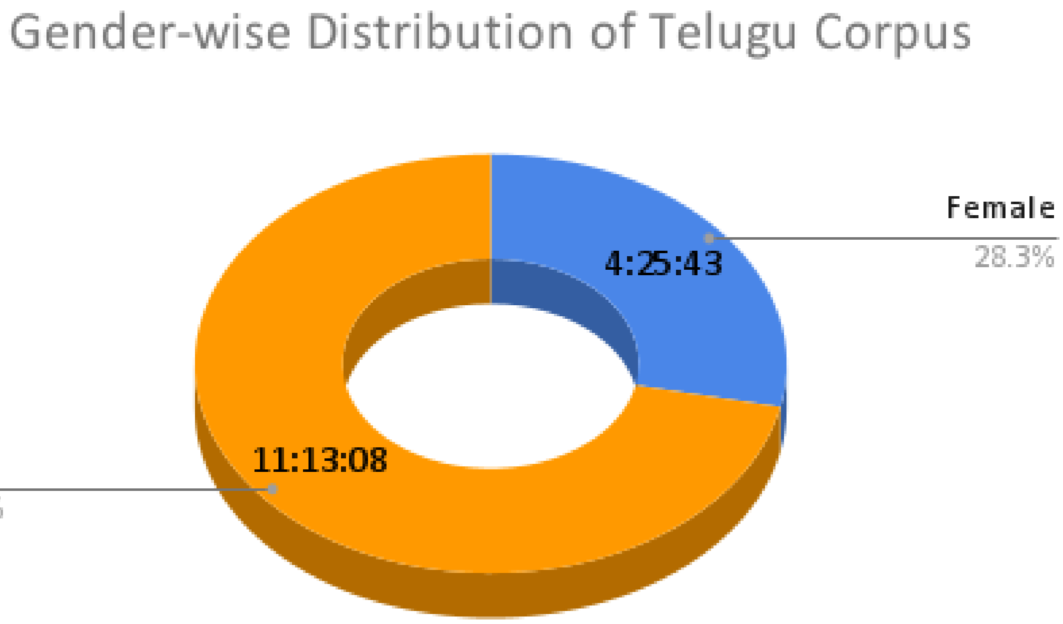

Figure 32: Gender-wise Distribution of Telugu Corpus

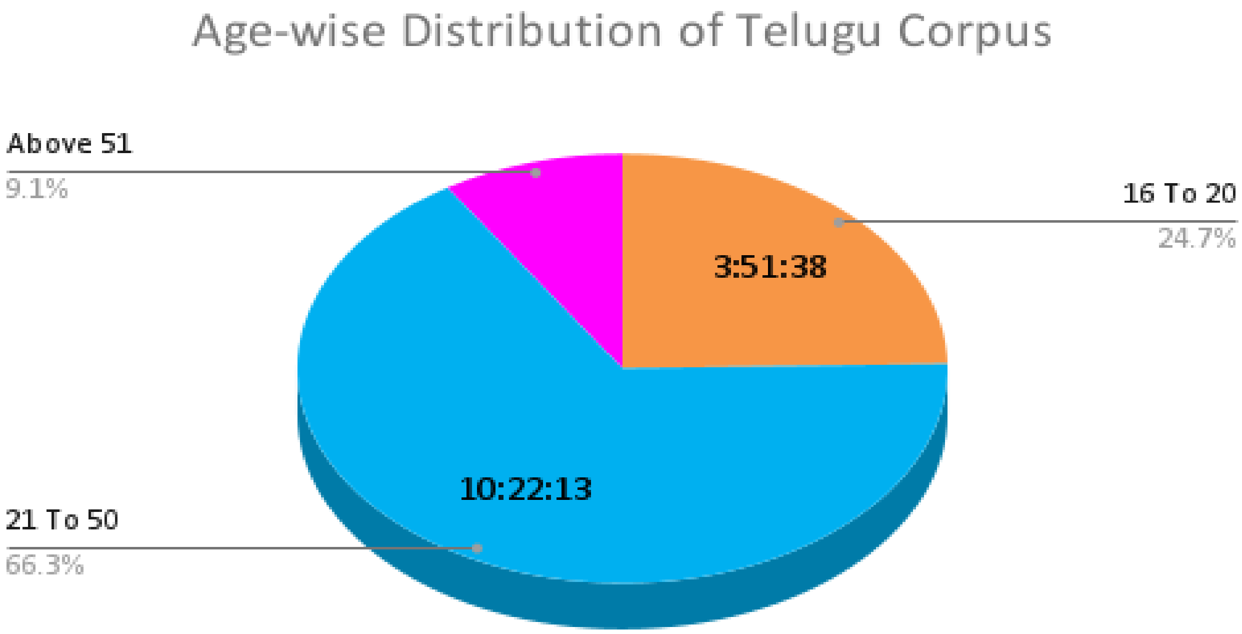

Figure 33: Age-wise Distribution of Telugu Corpus

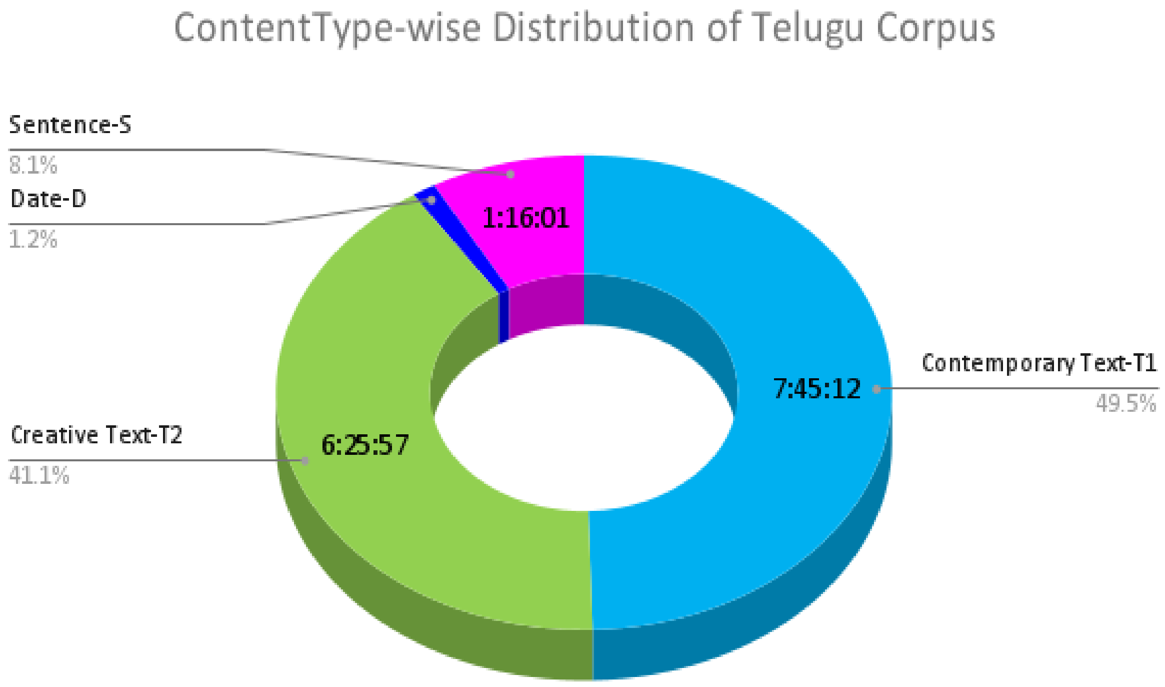

Figure 34: Content Type-wise Distribution of Telugu Corpus

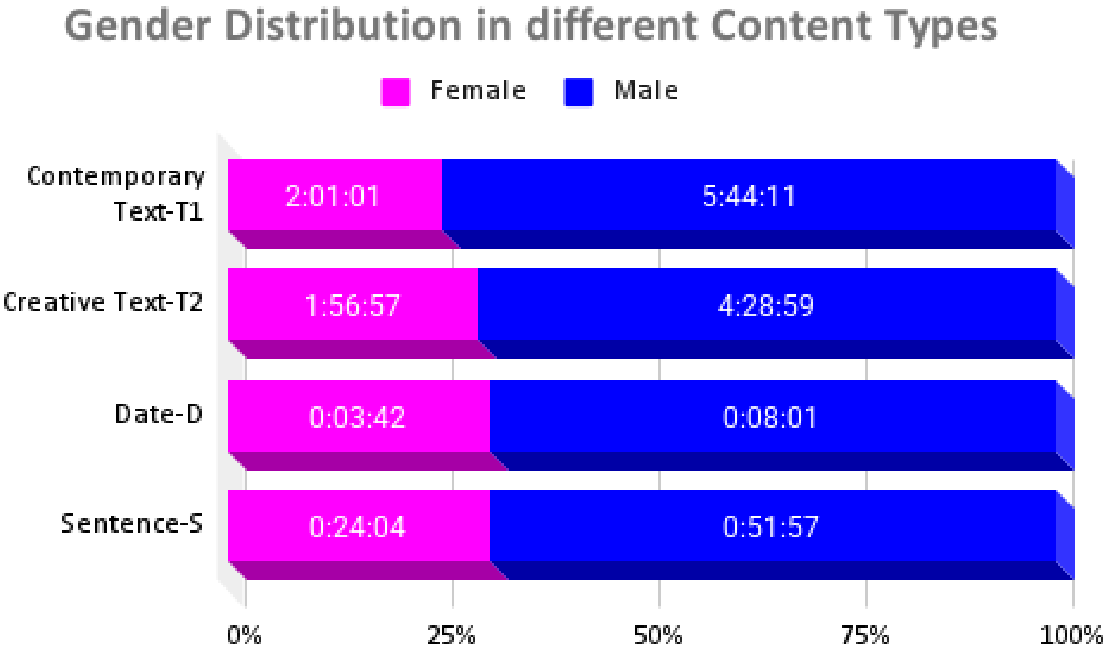

Figure 35: Gender Distribution in different Content Types of Telugu Corpus

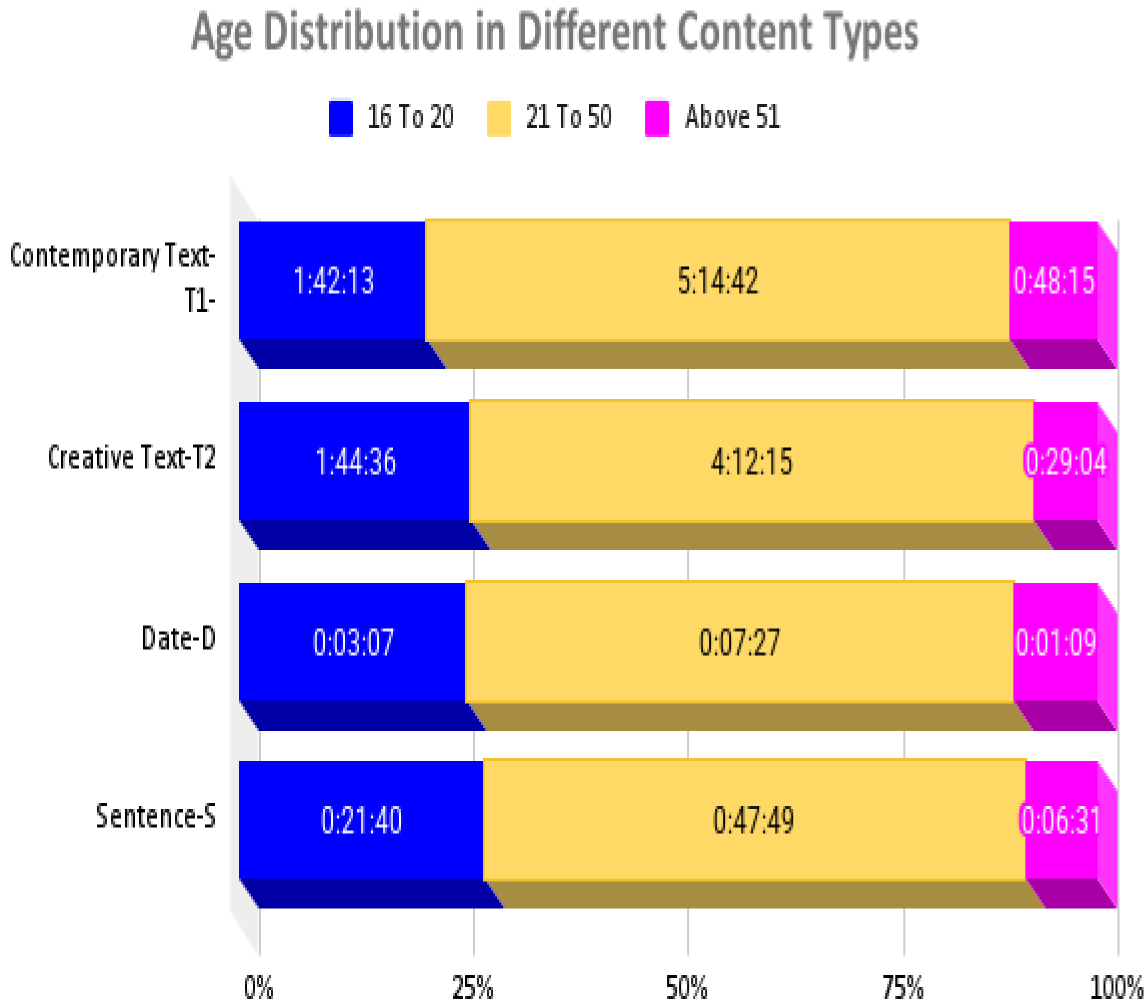

Figure 36: Gender Age Distribution in different Content Types of Telugu Corpus

13.3.1 DURATION OF TELUGU SENTENCE ALIGNED SPEECH DATA

The total duration of Telugu Sentence Aligned Speech Corpus is 15:38:53 (hh:mm:ss) comprising 9,548 audio segments from 80 speakers. The table below shows the duration of each of the content types and their distribution across a few factors in Telugu Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:10:06.956511	02:01:00.580352	07:45:11.782557
		21To50	00:19:44.900154		
		Above51	00:31:08.723688		
	Male	16To20	00:32:06.363889	05:44:11.202204	
		21To50	04:54:57.736923		
		Above51	00:17:07.101392		
Creative Text-T2	Female	16To20	01:06:46.599404	01:56:57.106580	06:25:56.592549
		21To50	00:26:48.894143		
		Above51	00:23:21.613032		
	Male	16To20	00:37:50.370389	04:28:59.485969	
		21To50	03:45:26.171128		
		Above51	00:05:42.944451		
Date-D	Female	16To20	00:02:13.107119	00:03:42.200776	00:11:43.549917
		21To50	00:00:35.817095		
		Above51	00:00:53.276562		
	Male	16To20	00:00:54.377334	00:08:01.349140	
		21To50	00:06:50.710973		
		Above51	00:00:16.260834		
Sentence-S	Female	16To20	00:15:00.536508	00:24:04.015859	01:16:00.798041
		21To50	00:04:18.311267		
		Above51	00:04:45.168084		
	Male	16To20	00:06:39.692053	00:51:56.782182	
		21To50	00:43:31.162670		
		Above51	00:01:45.927459		

Table 26: Representation of Telugu Sentence Aligned Speech Data Duration

13.4 SUMMARY OF SPEAKERS

The table below shows the total number of speakers and their distribution in the Telugu Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	15	7	22
21To50	4	47	51
Above51	5	2	7
Total	24	56	80

Table 27: Distribution of Speakers of Telugu Sentence Aligned Speech Data

13.5References

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
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, Kavitha Lenin, Rajesha N., Manasa G., 2019. Telugu Raw Speech Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Thirupal C Reddy & Gangaraju H. 2019. A Gold Standard Telugu Raw Text Corpus. Central Institute of Indian Languages, Mysore.

14 TEXT TO SPEECH CORPUS

Narayan Kumar Choudhary

14.1 INTRODUCTION

The Linguistic Data Consortium for Indian Languages (LDC-IL), established by the Ministry of Education, Government of India, serves as a national repository of linguistic data at the Central Institute of Indian Languages, Mysore. It develops and distributes qualitative linguistic resources and software technologies to support and enhance Indian languages. Details of previously published linguistic datasets by LDC-IL are available in [1], [2] and [3].

A text-to-speech (TTS) is a speech synthesis technology that converts written text to a natural-sounding speech. Text-to-Speech (TTS) applications utilize speech synthesis technology to convert textual content into audible speech, making them a valuable tool across various domains. These applications analyze text to determine proper pronunciation and generate corresponding speech output. By enabling users to listen to digital content, TTS technology enhances accessibility and usability. It is widely utilized in navigation systems, customer service interfaces, and assistive technologies, particularly benefiting individuals with visual impairments by providing digital accessibility.

LDC-IL developed a high quality Text-to-Speech (TTS) dataset contains speech recordings and the corresponding transcriptions along with the metadata. The dataset can be used in research, development, and evaluation of Text-to-Speech systems.

14.2 TTS PROMPT PREPARATION

The text prompts were majorly prepared from LDC-IL text corpus and ensured it covered a wide range of phonetic and prosodic features. The corpus includes all kinds of sentences with varying stress and pitch patterns. It is ensured that, no word of any sentence has more than 4 syllables. Sentences that have an overlap of more than 50% words are avoided.

14.3 DATA PREPARATION

The prepared prompts were recorded in studio environment by professional voice artists. LDC-IL preferred one male and female from 21-50 age groups. The specification of audio is as follows:

a.	**Sampling Rate**	:	48 kHz
b.	**Bit Rate**	:	16000 Mbps
c.	**Channels**	:	Stereo
d.	**File Format**	:	WAV

14.4 QUALITY ASSURANCE

Once the data is recorded, it undergoes transcription and subsequent evaluation by a third-party evaluator. The evaluator is responsible for verifying the transcription against the corresponding audio to identify and correct any inaccuracies. It is ensured that the mistakes encountered by a third party evaluator are also corrected by arbitrating it further by a third linguist.

14.5 REFERENCES

1. Choudhary, Narayan (ed.). 2019. Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-295-8.
2. Choudhary, Narayan (ed.). 2021. Compendium of Linguistic Resources in Indian Languages. Central Institute of Indian Languages, Mysore. ISBN: 978-81-948885-6-7.
3. Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2023. Compendium of LDC-IL Sentence Aligned Speech Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-19411-34-4.

15 Assamese Text to Speech Corpus

Syeda Mustafiza Tamim, Narayan Kumar Choudhary

15.1 Introduction

Assamese, which is primarily spoken in Assam and parts of Northeast India. Assamese has a rich literary heritage and a unique phonetic structure derived from the Eastern Nagari script [1]. According to some recent linguistic studies, the two dialectal groups can be further divided into four different varieties namely Eastern also known as Xiboxagoria, Central, Kamrupi and Goalparia group. Here, the Central group is spoken in Nagaon, Sonitpur, Morigaon and other adjoining areas; the Kamrupi variety is spoken in Kamrup, Barpeta, Nalbari, etc.; and the Goalparia group is spoken the Goalpara region. [2] and [3] are the datasets available for Assamese. Assamese Text to Speech Corpus provides a high-quality and phonetically balanced dataset of Assamese speech recordings along with their corresponding transcriptions. It is specifically designed for training, testing, and evaluating Assamese TTS models while also facilitating linguistic analysis of the language's phonetic, prosodic, and intonational characteristics.

15.2 Overview

The corpus includes assertive (statements), interrogative (questions), imperative (commands), and exclamatory sentences, as well as sentences with different stress and pitch patterns. The text prompts were majorly prepared from [3]. A few sentences were manually added to ensure the command and exclamation words. Assamese TTS corpus has 44:46:25 duration of data.

LDCIL follows a standard naming convention for storing audio files. Each audio file name consists language notation, gender id, sentence id, and the audio format name.
A typical audio file name is given below:
AS-TTS-F01-S-000001.wav
AS-TTS-M01-S-000001.wav

The summary of the Assamese TTS corpus is as follows:

Category	Sentence Count	Word Count	Male (Duration)	Female (Duration)	Total Duration
Statement	11,304	96,549	16:14:25	16:49:18	33:03:43
Question	1,683	14,678	02:30:12	02:03:00	04:33:12
Command	1,598	9,824	01:50:18	01:29:18	03:19:36
Exclamation	1,697	10,704	01:57:59	01:51:53	03:49:52
Total	16,282	1,31,755	22:32:55	22:13:30	44:46:25

Table 28: Summary of the Assamese TTS Corpus

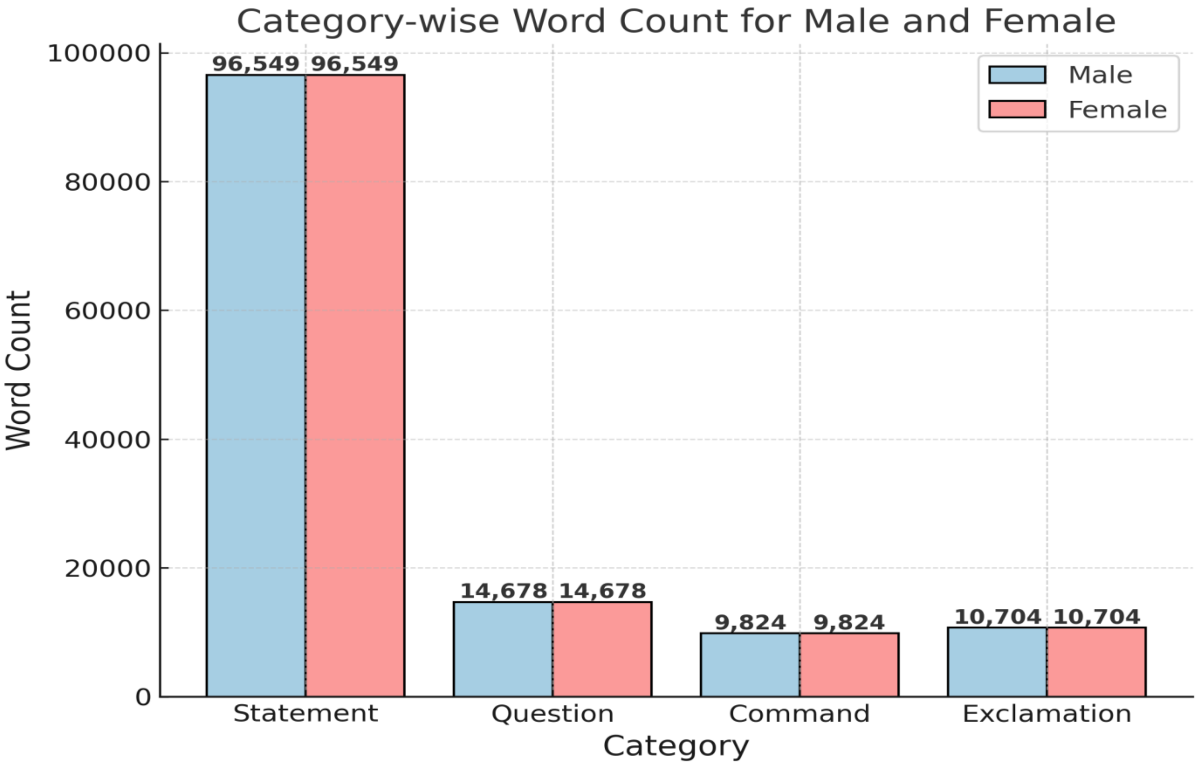

Figure 37: Gender-Wise Word Count Category Chart – Assamese Text-to-Speech Corpus

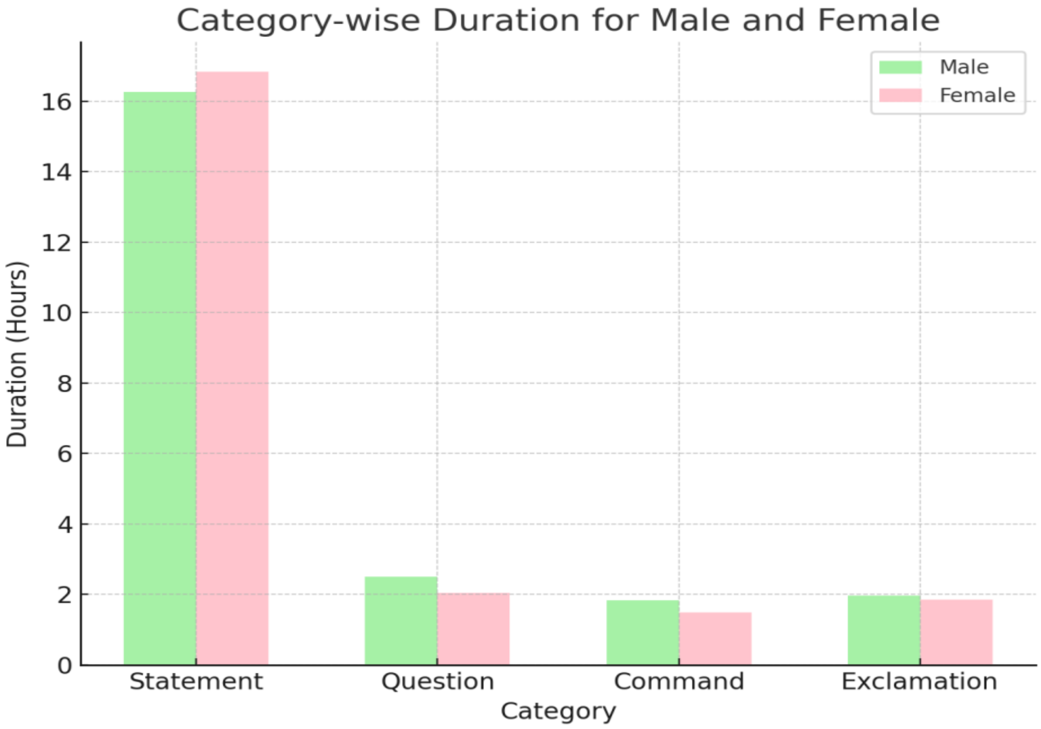

Figure 38: Gender-wise Duration Category Chart – Assamese Text-to-Speech Corpus

15.3 REFERENCES

1. Tamim, S. M., Adhyapak, P., Rajesha, N., Manasa, G., Srikanth, D., Choudhary, N. K., & Mohan, S. (2023). Assamese Sentence Aligned Speech Corpus. Central Institute of Indian Languages. e-ISBN: 978-81-948885-4-3.
2. Ramamoorthy, L., Choudhary, N. K., Sharma, A., Gogoi, A., Kalita, J., Bharadwaj, S., Bora, P., Adhyapak, P., Tamim, M., Rajesha, N., & Manasa, G. (2021). Assamese Raw Speech Corpus. Central Institute of Indian Languages. e-ISBN: 978-81-948885-5-0.
3. Hussain, T., Adhyapak, P., Tamim, S. M., Rajesha, N., Manasa, G., Choudhary, N., & Ramamoorthy, L. (n.d.). Assamese Raw Text Corpus. Linguistic Data Consortium for Indian Languages, Central Institute of Indian Languages

16 MAITHILI TEXT TO SPEECH CORPUS

Shantanu Kumar, Narayan Kumar Choudhary

16.1 INTRODUCTION

Maithili is a language spoken in India and Nepal and has six major varieties: Standard Maithili, Southern Standard Maithili, Western Maithili, Eastern Maithili, Chhika-chhiki, and Jolha Boli [1]. Based on the geographical area provided in his work covering the varieties, four major varieties, namely, the Bajjika, Sotipura, Angika, and Surjapuri varieties of modern Maithili correspond to Grierson's Western, Standard, Chhika-Chhiki, and Eastern varieties [2] [3]. [4] is the dataset available for Maithili. Maithili Text to Speech Corpus provides a high-quality and phonetically balanced dataset of Maithili speech recordings along with their corresponding transcriptions. It is specifically designed for training, testing, and evaluating Maithili TTS models while also facilitating linguistic analysis of the language's phonetic, prosodic, and intonational characteristics.

16.2 OVERVIEW

The corpus includes assertive (statements), interrogative (questions), imperative (commands), and exclamatory sentences, as well as sentences with different stress and pitch patterns. The text prompts were majorly prepared from [4]. A few sentences were manually added to ensure the command and exclamation words. Maithili TTS corpus has 30:28:43 duration of data.

LDCIL follows a standard naming convention for storing audio files. Each audio file name consists language notation, gender id, sentence id, and the audio format name.
A typical audio file name is given below:
MT-TTS-M01-S-000001.wav
MT-TTS-F01-S-000001.wav

The summary of the Maithili TTS corpus is as follows:

Category	Sentence Count (In Prompt)	Total Word Count (In Prompt)	Male (Duration)	Female (Duration)	Total Duration
Statement	11,551	80,341	09:37:38	11:21:47	20:59:25
Question	1,500	15,368	01:32:08	01:57:40	03:29:48
Command	1,586	7,946	00:57:56	01:06:34	02:04:30
Exclamation	1,493	15,822	01:35:49	02:19:12	03:55:01
Total	16,130	1,19,477	13:43:31	16:45:12	30:28:43

Table 29: Summary of the Maithili TTS Corpus

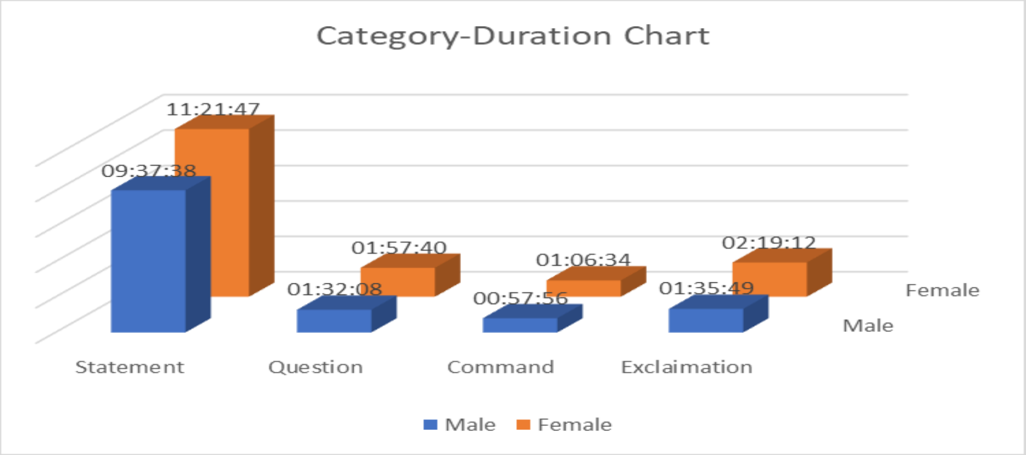

Figure 39: Gender-Sentence Category Chart – Maithili Text-to-Speech Corpus

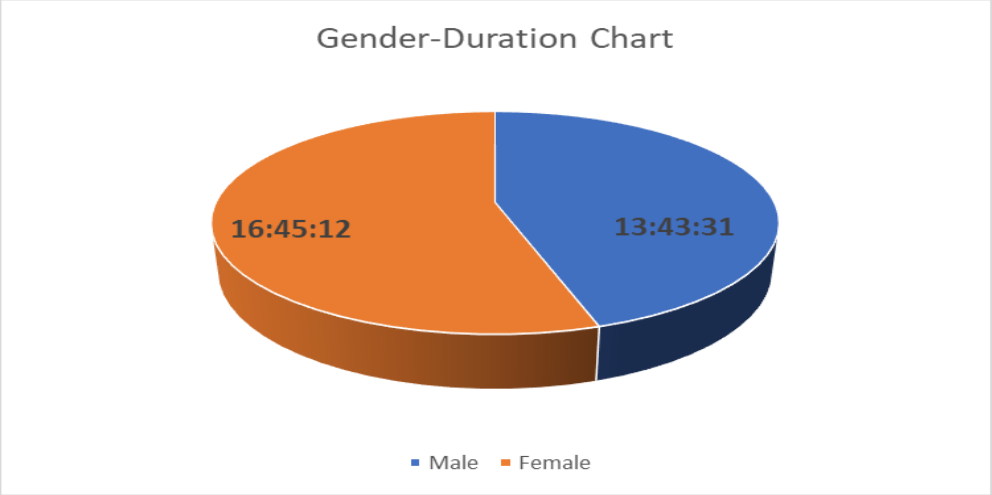

Figure 40: Gender-Duration Chart – Maithili Text-to-Speech Corpus

16.3 REFERENCES

1. Grierson, George Abraham. Linguistic survey of India. Vol. 1. Office of the superintendent of government printing, India, 1928.
2. Jha, Subhadra. "The formation of the Maithili language." PhD diss., Luzac London, 1958.
3. Jha, Govinda. Uchchatar Máithilī Vyākaran. Patna: Maithili Akadami, 1979.
4. Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-236-1.