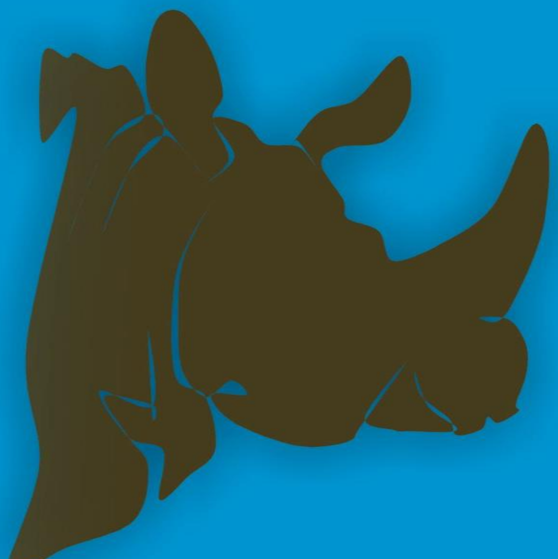
# আদৰ্শ অসমীয়া পাঠ্য ভাষাংশ



# A Gold Standard Assamese Raw Text Corpus

# Assamese Raw Text Corpus

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

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

CENTRAL INSTITUTE OF INDIAN LANGUAGES

Manasagangotri, Mysuru, Karnataka, India, 570006

www.ciil.org

Title: A Gold Standard Assamese Raw Text Corpus

*Authors*: Ramamoorthy L., Narayan Kumar Choudhary, Atreyee Sharma, Jahnobi Kalita, Samhita Bharadwaj, Taznin Hussain, Priyanshe Adhyapak, Mustafiza Tamim, Rajesha N., Manasa. G.

*e-ISBN: 978-81-948885-4-3*
*CIIL Publication No.: 1272*

*First published: AD 2021 May*
*Vaisakha 1943 Saka*

*Publisher*: C.G. Venkatesha Murthy, Director, CIIL

*Production Team*
*Head, Publication Unit: Umarani Pappuswamy*
*Officer-in-Charge, Publication Unit: Aleendra Brahma*
*Artist: H. Manohara*
*Staff-in-charge: R. Nandeesh*
*Compositor: M.N. Chandrashekar*
*Cover design: Rupesh Rai*

Contents

# Table of Figures

# 1 LDC-IL RAW TEXT CORPORA; AN OVERVIEW

This is a generic documentation of the LDC-IL raw text corpus which applies to all the languages covered in LDC-IL unless otherwise specified. However, this does not give the specifics of a language dataset.

The objective of language technology is to utilize the facilities of computer, to scientifically analyze language for retrieving verifiable proofs about properties of a language that enable the understanding of multi-dimensional nature of a language. Corpus of a language reflects the nature of the language. The larger and the more representative a corpus, the better it shows its nature.

A corpus is a large collection of language manifestation duly representing its aspects, mainly in text or spoken form. In case of sign language it is the collection of signs in visual form. The electronic text corpus is a collection of pieces of language text in electronic form, selected in accordance with the external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. Corpora are one of the major resources for language technology. Computers offer advantages like searching, selecting, sorting and formatting, which eases the language studies. Computers can avoid human bias in an analysis, thus making the result more reliable. Corpus serves as the basis for a number of research tasks within the field of Corpus Linguistics. It is the main resource for many modules of various applications like grammar checkers, spell checkers used in word editors etc. Indian languages often pose difficult challenges for developer community in Natural Language Processing/Artificial Intelligence. The technology developers building mass-application tools/products have for long been calling for availability of linguistic data on a large scale. However, the data should be collected, organized and stored in a manner that suits different groups of technology developers.

Over the years, a lot of efforts have been made to develop text corpora in Indian languages and several agencies have made contributed towards this including the government organizations, academic institutions as well as private bodies. However, the constant greed of more and more electronic data as required by the contemporary machine learning oriented technology models have proved that the data is still not sufficient for all the scheduled languages of India.

Linguistic Data Consortium for Indian Languages (LDC-IL) is one of the Government of India initiatives to develop linguistic corpora in Indian languages. Approved as a scheme in 2007 by the Ministry of Human Resource & Development, Government of India, LDC-IL started functioning at Central Institute of Indian Languages (CIIL), Mysore from April 15, 2008 when human resources got recruited for this scheme. The mission statement for this project is to develop “***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***[1]***.***"

The text datasets created under the LDC-IL ambit strives to fill the gap and provide more and more of electronic data for the NLP and language technology community such that the Indian languages get a boost and more of IT applications are available in these languages.

[1] Extract from the *Detailed Project Report* of LDC-IL.

# 2 LDC-IL APPROACH OF SAMPLING

Developing a written text corpus involves various factors like size of corpus, representativeness, quality of the text, determination of target users, selection of time-span, selection of documents etc. The data for the LDC-IL corpus are collected from books of general interest, textbooks, magazines, newspapers and Government documents of the contemporary text. The data is collected in accordance with prior set of criteria and with the convenience of material such as availability, proper format etc.

As a corpus is supposed to be representative of the language, there is no need to collect all the text from a given book. The representativeness of the corpus depends on a range of different kinds of text categories included in the corpus. LDC-IL corpora try to cover a wide range of text categories that could be representative of the language or language variety under consideration. Corpus representativeness and balance is closely associated with sampling.

LDC-IL collected text corpus from different sources. They are mainly books, magazines, and newspapers. The books are from literature and knowledge text books, magazines and newspapers are web crawled, or keyed in text or both. The newspaper and magazines are great resource of words which are hard to find in books because of the scarcity of those domain specific books in Indian languages.

LDC-IL has different Sampling approach over while extracting text from these three sources.

## 2.1 SAMPLING APPROACH FOR BOOKS

The books were identified so that the representation of different domains can be catered. After identifying the books, the next step is to extract typically 10 pages of text from it. LDC-IL follows a sampling method to collect the pages from a book. For example, if the book has 100+ pages we collect every 10th page and if the book has 200+ pages we collect every 20th page of the book. If the selected page contains pictures, tables etc, then its next or previous page, which may have the text content, will be chosen for the corpus. Even though one may find rare cases where partial or whole book is selected for the corpus, since the total corpus is going to be very large, such rare cases may not have an impact on balance of corpus. While selecting the book, the LDC-IL's motive is to select from wide variety of domains so that corpus can cover large part of vocabulary and should not miss out certain domain specific words.

Other generic principles that have been normally followed in the sampling tasks across languages are as follows:

- Contents containing obnoxious or vulgar texts have been avoided.
- New editions of the old books having a writing style prior to 1990 were not preferred. Rarely we may have text extracts from such books published prior to 1990 to ensure that the writing style is contemporary.

- For all texts containing short stories, sampling has been made by considering the short stories as a single entity and not based on the whole book containing all the short stories i.e. each page starting with a new short story have been sampled instead of the usual sampling method based on page numbers of the book.
- The data sampling personnel carried the category and sub-category list for ready reference in the field.
- Text extracts containing poems and formulae have been avoided.
- Pages containing diagrams, tables or figures have been avoided.
- Books containing less than 50 pages are not part of sampling.
- Texts having very small font have been enlarged during photocopying to make it look like 10 to 12 font size.
- If the text contains content other than the intended language, those texts have been avoided if the other language content is longer than one sentence.

## 2.2 Sampling Approach for Magazines

In case of magazine texts are small and from different domains so the whole magazine is to be considered to be included in corpus discarding advertisements, image captions, and tables etc. Magazine corpus usually includes different types of texts like cookery, health, cinema, stories, contemporary articles, etc.

## 2.3 Sampling Approach for newspaper

The newspaper corpus is contemporary text in nature. The text may contain political news, editorials, sports news etc. The news data does not have literary flourish. The news stories are on many unfamiliar domains, religious ideas, scientific principles etc. that have to be conveyed to the common people. So, it is expected that the writers would have captured these domains in a simple and meaningful way. Such write-ups have proper usage of vocabulary, correct language structure and effective phraseology. The newspaper articles may use colloquial, non-standard terms or jargons to attract the readers. The words used need to be expressive and represents the feeling and attitude towards the events. To cover such nuance of the language the newspaper are sampled to be part of the text corpus.

The News items of the paper is sampled based on the domains, classifieds, very small news snippets were avoided. Usually much of the newspaper is keyed.

# 3 LDC-IL TEXT CORPUS CATEGORIZATION

The LDC-IL corpus shows how people naturally use the language and it does not give imaginary, idealized examples. To satisfy this requirements we needed large amount of data otherwise the frequent items will be from some specific vocabulary or a particular style. Quantitative data gives somewhat accurate results of what occurs frequently and what occurs rarely in the language.

Each text source of corpus is different from others in form, function, content and features. This gives room to classify corpora into different categories. LDC-IL maintains a standard list of categories for which the text is to be collected. LDC-IL Identifies six major categories namely ‘Aesthetics’, ‘Commerce’, ‘Mass Media’, ‘Official Document’, ‘Science and Technology’, ‘Social Sciences’. These categories are further classified into 128 minor categories or sub-categories to cover various domains.

## 3.1 AESTHETICS

The Aesthetics category is one of the largest contributors to the LDC-IL corpus. This category contains sub-domains from Literature and Fine-arts. The text extracts are from literary sources. It is used to capture literature terms. Aesthetics text is collected from books. The text is probably any standard text which is descriptive in nature. It exhibits the language style of a particular period from which the text is taken. It is an extract of creative writing. It is made up of stories based on fiction, essays on various topics etc. These write-ups are mostly self-expressions of the writer. It captures the flow of language of the writer of the literary text.

The subdomains that are identified for mark-up in corpus under the Aesthetics are given below:

| Aesthetics | | | | |
|---|---|---|---|---|
| Fine Arts-Dance | Literary Texts | Literature-Novels | Autobiographies | Folk Tales |
| Fine Arts-Drawing | Literature-Criticism | Literature-Plays | Biographies | Folklore |
| Fine Arts-Hobbies | Literature-Diaries | Literature-Poetry | Cinema | Mythology |
| Fine Arts-Music | Literature-Essays | Literature-Epics | Culture | Photography |
| Fine Arts-Sculpture | Literature-Letters | Literature-Speeches | Handicrafts | Humour |
| Fine Arts-Musical Instruments | Literature-Children’s Literature | Literature-Text Books (School) | Literature-Travelogues | Literature-Science Fiction |
| | | | | Literature-Short Stories |

**Table 1: Subcategories of the Category Aesthetics**

## 3.2 COMMERCE

The trade is a part of the society. It exists and operates in association with various groups in society such as customer, suppliers, competitors, banks and financial institutions, Government agencies, trade unions. The trade domain has many domain specific words which need to be part of the corpus. The trade related books will bring such texts to the corpus.

The Subdomains that are identified for mark-up in corpus under the Commerce is given below:

| Commerce | | | | | |
|---|---|---|---|---|---|
| Industry | Accountancy | Share Market | Banking | Business | Career and Employment |
| Management | Finance | Tourism | | | |

**Table 2: Subcategories of the Category Commerce**

## 3.3 MASS MEDIA

Media is an integral part of everyday life for many people all over the world, at work and in the home. The text from this domain is contemporary in nature. The text may contain political news, editorials, or sports news. The major source of the Mass Media text category is newspaper; it contains words which are used in day-to-day life. Structurally, the language of mass media contains exposition, argument, description and narration. It includes different types of write up; consists of structures with different patterns, words and styles. All this is written in a language in which everyone can relate and understand. Some of the media prints are in the form of conversation or question answers. This data usually contains an interviewer and an interviewee. They usually consist dialogues. The interviewee may be a celebrity or a renowned personality from cinema, politics etc. The words used in such text are usually more personal and simple.

The Subdomains that are identified for mark-up in corpus under the Mass Media is given below:

| Mass Media | | | | | |
|---|---|---|---|---|---|
| Article | Classifieds | General News | Obituary | SMS | Religious/Spiritual News |
| Business News | Discussions | Interviews | Political | Social | Sports News |
| Cinema News | Editorial | Letters | Speeches | Weather | Health |

**Table 3: Subcategories of the Category Mass Media**

## 3.4 OFFICIAL DOCUMENT

The usage of language in official documents is highly standard, unambiguous, straight forward and structurally modified. The communication intended in official documents are intended about some action, or some enquiry or proceedings of some assemblies. This text usually is to get the due representation of such domain specific terminologies of administration, official document category is included.

The Subdomains that are identified for mark-up in corpus under the Official Document is given below:

| Official Document | | | |
|---|---|---|---|
| Administration | Legislature | Parliamentary/Assembly Debates | Police Documents |

**Table 4: Subcategories of the Category Official Documents**

## 3.5 SCIENCE AND TECHNOLOGY

The science and technology domain contains text extracts from various scientific books, articles of magazines, journals etc. These texts are also called as knowledge texts. The language structure and usage of words are different from the language of day-to-day life. The terminologies that are from this domain will have highest number of loan words because the subject in the text is usually global. To get the due representation of such domain specific terminologies, the Science and Technology category is included.

The Subdomains that are identified for mark-up in corpus under the Science and Technology is given below:

| Science and Technology | | | | | |
|---|---|---|---|---|---|
| Agriculture | Biotechnology | Engineering-Civil | Forestry | Medicine | Statistics |
| Architecture | Botany | Engineering-Electrical | Geology | Micro Biology | Astrology |
| Textile Technology | Educational Psychology | Engineering-Electronics Communication | Text Book (Science) | Computer Sciences | Language Technology |
| Chemistry | Naturopathy | Engineering-Mechanical | Horticulture | Oceanology | Veterinary |
| Ayurveda | Criminology | Engineering-Others | Astronomy | Physics | Film Technology |
| Bio Chemistry | Homeopathy | Environmental Science | Logic | Psychology | |
| Biology | Yoga | Engineering-Chemical | Mathematics | Sexology | Zoology |

**Table 5: Subcategories of the Category Science and Technology**

## 3.6 SOCIAL SCIENCES

Language is a medium for creation and maintenance of human society so language in social sciences category correlates the linguistic features of the dynamic society. Human development and reformation happening in different communal context hence all the social knowledge and reality could be reflected in this text category.

The Subdomains that are identified for mark-up in corpus under the Social Sciences is given below:

| Social Sciences | | | | | |
|---|---|---|---|---|---|
| Anthropology | Food and Wellness | Personality Development | Physical Education | Text Book (Social Science) | Philosophy |
| Archaeology | | | | | Journalism |
| Demography | Fisheries | Library Science | Law | Sports | Geography |
| Economics | | | | | Religion / Spiritual |
| Education | Home Science | Political Science | Public Administration | Health and Family Welfare | Sociology |
| Epigraphy | | | | | Linguistics |

**Table 6: Subcategories of the Category Social Sciences**

# 4 LDC-IL TEXT DATA ENCODING AND FORMAT

The collected data should be encoded in a machine readable form for further analysis. While storing the data one has to keep some standards so that the data is easy to store and retrieve in long term. The encoding being used in LDC-IL Text corpus is Unicode and stored in XML format. Large scale language resource depends on the metadata. Metadata is an authentic source to prove the quality of the data. Metadata should have the subject information, source information and encoding information.

The selected text along with metadata information is indexed with a five digit unique number to get keyed-in. Each text fragment of selected book is typed as corpus file with xml extension. The given unique Index number gets prefixed with the LDC-IL notations which make the file name of the XML file. Sometimes the XML file names carry small case alphabets enclosed in braces. This is done if the book title carries different type of textual topics, so that each chapter, in the selected book title which may be related to different topics, chapters etc., can be differentiated. This helps the text content get categories based on the context.

# 5 LDC-IL TEXT CORPUS METADATA

It is imperative to maintain metadata of the entire data collection for linguistic analysis. The collected data are arranged with its metadata information such as its category, subcategory, title of the text, author name, source, publisher name, year of publication, page numbers etc. This information helps the users to retrieve the data easily from the database/repository. Metadata gives authenticity to the text by way of providing the details of how the data was created in the first instance and what is its content about. The following table shows the legend used in the metadata and provides description of them.

| # | Legend | Description |
|---|---|---|
| 1 | Filename | Represented by "docID" tag in the XML files. This is a unique file number across the datasets. |
| 2 | Project Description | This gives a brief of the project under which the file was generated. As CIIL has been involved into corpus creation over a long period time, including before the inception of LDC-IL scheme, there might be some data for a few languages which might have come from different projects e.g. the CIIL Corpus or CIIL-KHS corpus. This field indicates the source of the project. |
| 3 | Sampling Description | This information is a verifiable proof for the corpus. It will have the information of selected page numbers of the book for corpus. |
| 4 | Category | Specifies the domain of the text. |
| 5 | Subcategory | Specifies the sub-domain of the text. |
| 6 | Text | Specifies the type of the source text i.e. whether its origin is a book, a magazine or a newspaper. |
| 7 | Title | Specifies the title of the source text. It contains mostly books but if magazines or newspapers occur, their respective are provided here. |
| 8 | Volume | Specifies volume number the title, if any. |
| 9 | Issue | Specifies issue number the title, if any. |
| 10 | Text Type | Is mostly blank however sometimes it is used to provide the broad topic of the news items e.g. whether it is a political news or editorial or sports news etc. |
| 11 | Headline | This information is a verifiable proof for the corpus. This is normally the heading of the chapter of the selected sample. Gives the fine tuned information of the topic present in particular file. |
| 12 | Author | Specifies the name of the author. |
| 13 | Editor | Specifies the name of the editor. |
| 14 | Translator | Specifies the name of the translator. |
| 15 | Words | Specifies the total number of words in the file. |
| 16 | Letters | Specifies the total number of UTF8 characters in the file. |
| 17 | Publishing Place | Specifies the place where the title was published. |
| 18 | Publisher | Specifies the name of the publisher. |
| 19 | Published Year | Specifies the publishing year. |
| 20 | Index | Is the index number or ID of the file. It is noted inside the XML file. It is mostly the same as the file name. |
| 21 | Date | Date when the file was digitized/inputted. |
| 22 | Input | Name of the Data Inputter, if the file has been typed. |
| 23 | Proof | Name of the Proof reader. |
| 24 | Language | Name of the language. |
| 25 | Script | Name of the script the text is written in. |

**Table 7: Metadata Legends for LDC-IL Text Data**

Typical Metadata Mark-ups in a text corpus file structure is given below.

| | | | |
|---|---|---|---|
| <?xml version="1.0" ?> | | | |
| <?xml-stylesheet type="text/css" href="home.css"?> | | | |
| <Doc id="mal-w-media- | ML00172 | " | lang="Malayalam"> |
| <Header type="text"> | | | |
| <encodingDesc> | | | |
| <projectDesc> | CIIL-Malayalam Corpora, Monolingual Written Text | | </projectDesc> |
| <samplingDesc> | Simple written text only has been transcribed. Diagrams, pictures and tables have been omitted. Samples taken from page 30-31,50-51,70-71,94-95,114-115,132-133,152-153,172-173,192-193,210-211 | | </samplingDesc> |
| </encodingDesc> | | | |
| <sourceDesc><br><biblStruct><br><source> | | | |
| | <category> | Aesthetics | </category> |
| | <subcategory> | Literature-Novel | </subcategory> |
| | <text> | Book | </text> |
| | <title> | Kalapam | </title> |
| | <vol> | | </vol> |
| | <issue> | | </issue> |
| </source> | | | |
| <textDes> | | | |
| | <type> | | </type> |
| | <headline> | | </headline> |
| | <author> | ShashiTharoor | </author> |
| | <editor> | | </editor> |
| | <translator> | Thomas George | </translator> |
| | <words> | 2745 | </words> |
| </textDes> | | | |
| <imprint> | | | |
| | <pubPlace> | India-Kottayam | </pubPlace> |
| | <publisher> | DC Books | </publisher> |
| | <pubDate> | 2006 | </pubDate> |
| </imprint> | | | |
| <idno type="CIIL code"> | | Kerala University Campus Library- 13535 | </idno> |
| <index> | | ML00172 | </index> |
| </biblStruct> </sourceDesc> | | | |
| <profileDesc> <creation> | | | |
| | <date> | 26-Apr-2010 | </date> |
| | <inputter> | Remya K | </inputter> |
| | <proof> | | </proof> |
| </creation> | | | |
| <langUsage> | | Malayalam | </langUsage> |
| <ScriptUsage> | | Malayalam | </ScriptUsage> |
| <wsdUsage> | | | |
| <writingSystem id="ISO/IEC 10646"> Universal Multiple-Octet Coded Character Set (UCS). </writingSystem> | | | |
| </wsdUsage> | | | |
| <textClass> | | | |
| <channel mode="w"> | | Print | </channel> |
| <domain type="public"> | | | </domain> |
| </textClass> </profileDesc> </Header> | | | |
| <text> <body> | | | |
| <p> | | | </p> |
| <p> | | | </p> |
| </text> </body> < /Doc> | | | |

# 6 LDC-IL Text Corpus and Naming conventions

The selected hardcopies were marked for sampling and given to typists by concerned language experts. LDC-IL has built an in-house corpus developing application and stores it in a repository database. The samples get typed in xml format through a software application built for it in LDC-IL. Each sampling is a corpus file and gets typed and saved in Unicode standards. Each corpus file has unique file name. One can say the corpus is indexed through file names. Typically each corpus file is an extract of a book of a particular title. The LDC-IL corpus file name follows certain naming convention. The naming convention is based on language and source of text. Every scheduled language has a notation for each kind of source of corpus. The notation is prefixed to a five digit number to create a unique corpus filename.

The LDC-IL notations for Indian Scheduled languages are given below.

| # | Language | ISO 639 Language Code | Script | Notation as per Source of Corpus | | | |
|---|---|---|---|---|---|---|---|
| | | | | Book | Magazine | News Paper | News Web |
| 1 | Assamese | asm | Assamese | AS | ASM | ASN | ASNW |
| 2 | Assamese | ben | Assamese | BE | BEM | BEN | BENW |
| 3 | Bodo | brx | Devanagari | BD | BDM | BDN | BDNW |
| 4 | Dogri | doi | Devanagari | DG | DGM | DGN | DGNW |
| 5 | Gujarati | guj | Gujarati | GJ | GJM | GJN | GJNW |
| 6 | Hindi | hin | Devanagari | HN | HNM | HNN | HNNW |
| 7 | Kannada | kan | Kannada | KA | KAM | KAN | KANW |
| 8 | Kashmiri | kas | Persio-Arabic | KS | KSM | KSN | KSNW |
| 9 | Konkani | kok | Devanagari | KO | KOM | KON | KONW |
| 10 | Maithili | mai | Devanagari | MT | MTM | MTN | MTNW |
| 11 | Malayalam | mal | Malayalam | ML | MLM | MLN | MLNW |
| 12 | Manipuri | mni | Assamese | MN | MNM | MNN | MNNW |
| 13 | Marathi | mar | Devanagari | MA | MAM | MAN | MANW |
| 14 | Nepali | nep | Devanagari | NP | NPM | NPN | NPNW |
| 15 | Odia | ori | Odia | OD | ODM | ODN | ODNW |
| 16 | Punjabi | pan | Gurmukhi | PN | PNM | PNN | PNNW |
| 17 | Sanskrit | san | Any Script | SA | SAM | SAN | SANW |
| 18 | Santali | sat | OlChiki | SN | SNM | SNN | SNNW |
| 19 | Sindhi | snd | Persio-Arabic | SI | SIM | SIN | SINW |
| 20 | Tamil | tam | Tamil | TA | TAM | TAN | TANW |
| 21 | Telugu | tel | Telugu | TE | TEM | TEN | TENW |
| 22 | Urdu | urd | Persio-Arabic | UR | URM | URN | URNW |

**Table 8: LDC-IL notations for Indian Scheduled languages**

Consider the example of Malayalam, The text taken from Malayalam book for LDC-IL Malayalam Text Corpus always starts with ‘ML’ followed by 5 digit numbers which is continuous, where as text collected from Malayalam Magazine starts with ‘MLM’ followed by 5 digit numbers. If the source is from Newspaper then ‘MLN’ notation will be followed where as if the News is taken from Web source ‘MLNW’ will be used as notation.

In certain cases, if the book is chaptered, the headline of each chapter changes, to capture the change of the topic. If the language experts wish to break the sampling of a book into different smaller files, then the file name will get attached with roman small letter suffixed and enclosed in braces.

Such file names could be ‘ML00001(a)’, ‘ML00001(b)’, ‘ML00001(c)’, ‘ML00001(d)’ etc.

# 7 PROOF READING

Once it is in digital form, the same is proof read so that it is free from any kind of typographical errors. Proofing is the next process of corpus building. Since the typed corpus may carry errors because of various reasons like speed of the typist and typist not belonging to the language community, the proofing is done by the language experts.

While proofing of a corpus file is done in LDC-IL, the following things are taken care of

1. Removing the poetic text, if any poem or poetic structure occurs within the running text
2. If there are incomplete sentences typed (generally at the end of the paragraph) the sentence is removed up to the logical ending of the previous sentence.
3. Verifying the difference between the visargaha and colon ':' symbol, and to ensure that the correct symbol/punctuation is used in the correct place.
4. During Content cleaning focus stays on the corrections of typographical errors and spacing. If there is a space preceding a punctuation mark, space is removed, unless it is there in the actual text itself (i.e. hard copy of the text).
5. If there is any mismatch between the hard copy and the input corpus file, it is ensured that the corpus file should be faithful to hard copy.
6. It is ensured that the Title, Author, Headline fields of the XML files is written in Roman using the LDC-IL transliteration scheme. The LDC-IL Transliteration scheme can be referred on the LDC-IL website. Also, the LDC-IL transliteration tool from Roman to Indian Scripts and vice versa is available for download on the LDC-IL website.
   Link to download LDC-IL Transliteration Scheme:
   http://ldcil.org/Tools/CorporaToolsPackage/LDC-IL%20Transliteration%20Scheme.pdf

   Link to download the LDC-IL Transliteration Tool (.exe file):
   http://ldcil.org/Tools/LDC-IL%20Transliterator.zip

Proof reading is used to correct clear cases of spelling mistakes, splitting sentences or words, removing unnecessary repeated paragraphs, sentences, phrases, words. Moreover, it includes removing unwanted texts from the corpus such as foreign script sentences and incorrect use of ungrammatical sentences.

# 8 COPYRIGHT

Anyone intending to put together a corpus for commercial purposes must always obtain the permission from the publishers of the source texts. Many commercially available corpora contain texts from a large number of sources and obtaining permission to use these can be a very cumbersome and financially costly process. However, LDC-IL took up the task and managed to get the consent of most of the copyright holders or has at least communicated to them that the text extracts from their sources are being used in the language sampling task which may also be used commercially.

Considering LDC-IL is a government initiative taken up in the larger public interest and the corpus is used for the development of language, most of the publishers and authors generously agreed to archive the samples of their text materials in corpus. Some of the authors even suggested and offered their other content which are not yet part of the LDC-IL corpus. Government publishers too expressed no objections regarding since LDC-IL itself is an initiative of Govt. of India. Private publishers also gave permission considering that LDC-IL is only using a part of a text, and it will not harm their business anyway. LDC-IL thanks all of them for the co-operation.

For some of the content where we have not yet got the explicit consent of the copyright holders, we have sent them the letters asking for the same. If any of the copyright holders disagree to consent, they may write so to us and their respective text will be removed from the sampling corpus and the same will be intimated to all the license holders of the respective dataset and they will have to abide by it.

# 9 ASSAMESE RAW TEXT CORPUS

## 9.1 INTRODUCTION

Assamese or Oxomiya is the language spoken by the natives of the state of Assam in Northeast India. It is also the official language of Assam. It is spoken in some parts of Arunachal Pradesh, Nagaland and in other Northeast Indian states. However, small pockets of Assamese speakers can also be found in Bhutan and Bangladesh. The origin and growth of the Assamese language is not simple and clear. Some writers think that its source is to be found in the Sanskrit or Vedic literature. But, Assamese, a branch of Indo-Aryan along with the cognate languages, Maithili, Bengali and Oriya, developed from Magahi Prakrit. According to linguist Suniti Kumar Chatterji, the Magahi Prakrit in the east gave rise to four Apabhramsa dialects namely-Radha, Vanga, Varendra and Kamarupa. The Kamarupa Apabhramsa gave rise to the Assamese language in the Brahmaputra valley. Though early compositions in Assamese exist from the 13th century yet the earliest relics of the language can be found in paleographic records of the Kamarupa Kingdom from the 5th century to the 12th century. Some compositions also date to the 14th century, during the reign of the Kamata king Durlabh Narayana of the Khen dynasty, when Madhav Kandali composed the Kotha Ramayana.

Assamese text corpus is collected from various libraries in Assam mostly from Guwahati and Jorhat. The greater part of the text has been taken from Guwahati District Library and Jorhat District Library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of text but some categories like physics, chemistry, economics have very less amount of text. Literary texts are easily available in Assamese but getting scientific text is exceedingly difficult. In some categories, like epigraphy, finance, oceanology, texts are too rare to find in Assamese.

## 9.2 PERIOD OF ASSAMESE TEXT INCLUDED IN THIS CORPUS

While modern assamese language has a continuum of history starting from 14th century, for the purpose of this corpus, we have focused on modern Assamese writing and no text written prior to 1990 is part of it.

## 9.3 PECULIARITIES OF ASSAMESE LANGUAGE

The Corpus of Assamese text can be broadly classified into two: literary text and non-literary text. These two explicitly show their differences in terms of frequency of word usage and variety that it brings into corpus. Literary texts are texts in narrative style, and they also contain elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are text whose primary purpose is to convey information. Examples of non-literary texts are text about various scientific or technical subjects, legal documents, articles in academic journals. In literary text, language has emotional elements, cultural information, dialectical variations, ambiguity etc. But technical or scientific terms, foreign words etc. have widely appeared in non-literary texts.

Some of salient features of Assamese language are discussed below.

The Assamese phonemic inventory consists of eight vowels, ten diphthongs and twenty consonants. The velar fricative /x/ is the hallmark of this language. Assamese language is rich in consonant clusters.

Eastern Indic languages like Assamese, Bengali, Sylheti, and Oriya do not have a vowel length distinction rather it has a wide set of back rounded vowels. In case of Assamese there are four back rounded vowels.

Gender in Assamese is not grammatically marked. Another important feature of this language is that the verbs in Assamese are negativized by adding /-n/ before the verb with /n/ picking up the initial vowel of the verb.

For example, *lage* which means “want” and to negativize it a negative particle /-na/ is added to the verb like *nalage* meaning “does not want”. Assamese has a huge collection of classifiers which are extensively used for different objects. The basic word order in Assamese language is SOV. However, it normally allows scrambling, that is, Assamese, to a certain extent has a flexible word order system.

## 9.4 ASSAMESE SCRIPT

The Assamese script which is also known as Oxomiya Akhor or Oxomiya Lipi is a variant of the Eastern Nagari script and is also used for Bengali and Bishnupriya Manipuri. In Unicode, it may be referred to as Bangla or Bengali script. The Eastern Nagari script belongs to the Brahmic family of scripts and has a continuous history of development from Nagari script, a precursor of Devanagari.

## 9.5 PRINCIPLES OF DATA SAMPLING

Assamese text data sampling strictly followed the generic guidelines of LDCIL text corpus collection which are noted in the generic LDC-IL corpus documentation.

## 9.6 FIELD WORKS UNDERTAKEN

Assamese text corpus is collected from various libraries in Assam, mostly from Guwahati. The text materials were collected by conducting six field works in the period of 2008 to 2012. The greater part of the text has been taken from different library of Guwahati and Jorhat.

Overall, the following libraries served as the source of the Assamese text corpus:

- J B College, Jorhat.
- Assam Agricultural University, Jorhat.
- District Library, Jorhat.
- Personal Libraries, Jorhat.
- District Library, Guwahati.
- NERLC Library, Guwahati.
- Cotton College Library, Guwahati.
- Personal Libraries, Guwahati.

Collected text materials have been published at various places within Assam and other states of India, including Delhi.

Collecting text data from the field is a difficult job. Most of the libraries do not allow taking huge amount of text from their shelves at a time because it is against their rules and principles. For a particular period, they issue maximum three or four books. Even if the librarian allowed taking many books at a time, the photocopy kiosk had issues as there was a long queue.

Sometime photocopy attendants refused to photocopy randomly selected pages because of the long queue waiting and it took more time for them to turn the pages compared to continuous page photocopying that they are accustomed to. The field worker/linguist had to carry a huge list of photocopy bundles with them, which was many times burdensome to travel with.

Despite all the issues mentioned above, the linguists working on the data collection had to deal with and get going.

## 9.7 DATA INPUTTING

All the texts have been typed in Unicode using the InScript Keyboard directly onto the XML files. The data sets have been inputted by Ms. Roopa Devi, Rina Sarkar and Golap Borah who are the native speakers of Assamese.

## 9.8 VALIDATION AND NORMALIZATION WORKSHOPS

A 10-day workshop was conducted at Linguistic Data Consortium for Indian Languages from 24-June-2010 to 05-July-2010 with Ms. Karabi Gogoi (RP), Ms. Rita Sharma (RP), Sewali Deka (RP) and Pikumoni Chutia (RP) as experts. The experts suggested that the Assamese text corpus should remain true to the text and typo errors have been corrected.

## 9.9 PROOFREADING

Assamese text data has been proofread by internal Resource Persons such as Atreyee Sharma, Ashmrita Gogoi, Jahnobi Kalita, Muslima Begum, Chuchen Dutta, Amrit Upadhyay, Taznin Hussain, Priyanshe Adhyapak, Anupama Rabha, Arpan Jyoti Gogoi, Karishma Hazarika, Bidyut Bezbaruah, Bijoy Krishna Doley, Hemanta Konch, Rini Dehingia, Sharmistha Saikia, Jitu Borah, Rehna Sultana, Tulika Sarmah.

The collected text materials are contemporary and mainly published after 1990.

# 10 TRANSLITERATIONS IN LDC-IL ASSAMESE TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Assamese to Roman letters. Numeric characters were transliterated from Assamese to Hindu-Arabic system.

The LDC-IL transliteration scheme of Assamese to Roman is given below

LDC-IL Transliteration Schema
Assamese characters to Roman and Assamese Numerals to Hindu-Arabic

| Vowels and Vowel Signs | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| অ | আ | ই | ঈ | উ | ঊ | ঋ | এ | ঐ | ও | ঔ |
| | ◌া | ি◌ | ◌ী | ◌ু | ◌ূ | ◌ৃ | ে◌ | ৈ◌ | ে◌া | ে◌ৗ |
| a | A | i | I | u | U | x | E | ai | O | au |

| Consonants | | | | |
|---|---|---|---|---|
| ক | খ | গ | ঘ | ঙ |
| ka | kha | ga | gha | ng'a |
| চ | ছ | জ | ঝ | ঞ |
| ca | cha | ja | jha | nj'a |
| ট | ঠ | ড | ঢ | ণ |
| Ta | Tha | Da | Dha | Na |
| ত | থ | দ | ধ | ন |
| ta | tha | da | dha | na |
| প | ফ | ব | ভ | ম |
| pa | pha | ba | bha | ma |

| Symbols | | |
|---|---|---|
| ং | ঃ | ◌ঁ |
| M | H | m' |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| য | ৰ | ল | ৱ | শ | স | ষ | হ | ড় | ঢ় | য় | ৎ |
| ya | ra | La | wa | sha | sa | Sa | ha | D'a | Dh'a | Ya | t |

| Numerals(Bengali to Hindu-Arabic) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ০ | ১ | ২ | ৩ | ৪ | ৫ | ৬ | ৭ | ৮ | ৯ |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

# 11 OVERVIEW OF REPRESENTED DOMAINS

At LDC-IL, the Assamese text corpus size is: 1,01,27,030 Words drawn from 1,084 different titles. The total Corpus character size is 6,39,50,126 The data can be categorized into two classes: typed+cleaned. The representation of six major domains covered and has been shown in the table below:

| **Domain** | **Domain Word Count** | **Percentage** |
|---|---:|---:|
| Aesthetics | 5233452 | 51.68% |
| Commerce | 66924 | 0.66% |
| Mass Media | 3354996 | 33.13% |
| Official Document | 1298 | 0.01% |
| Science and Technology | 372790 | 3.68% |
| Social Sciences | 1097570 | 10.84% |
| **Total** | 10127030 | 100.00% |

**Table 9: Representation of the various domains in Assamese text corpus**

Each domain has several sub-domains, the following table shows the representation of main and sub-domains:

## 11.1 AESTHETICS

The Aesthetic domain of Assamese text corpus covers 30 subdomains bearing a total of 52,33,452 words along with the overall percentage of 51.68% The representational details are given in the table below:

| # | Subdomain | Word Count | Percentage (within Subdomain) | Overall Percentage |
|---|---|---|---|---|
| 1 | Autobiographies | 90101 | 1.72% | 0.89% |
| 2 | Biographies | 260592 | 4.98% | 2.57% |
| 3 | Cinema | 70760 | 1.35% | 0.70% |
| 4 | Culture | 220442 | 4.21% | 2.18% |
| 5 | Fine Arts-Dance | 80665 | 1.54% | 0.80% |
| 6 | Fine Arts-Drawing | 8119 | 0.16% | 0.08% |
| 7 | Fine Arts-Hobbies | 12749 | 0.24% | 0.13% |
| 8 | Fine Arts-Music | 53977 | 1.03% | 0.53% |
| 9 | Fine Arts-Musical Instruments | 190 | 0.00% | 0.00% |
| 10 | Fine Arts-Sculpture | 26746 | 0.51% | 0.26% |
| 11 | Folk Tales | 4560 | 0.09% | 0.05% |
| 12 | Folklore | 48052 | 0.92% | 0.47% |
| 13 | Handicrafts | 1697 | 0.03% | 0.02% |
| 14 | Humour | 29191 | 0.56% | 0.29% |
| 15 | Literary Texts | 1074762 | 20.54% | 10.61% |
| 16 | Literature-Children's Literature | 152813 | 2.92% | 1.51% |
| 17 | Literature-Criticism | 459582 | 8.78% | 4.54% |
| 18 | Literature-Diaries | 26657 | 0.51% | 0.26% |
| 19 | Literature-Epics | 14702 | 0.28% | 0.15% |
| 20 | Literature-Essays | 32880 | 0.63% | 0.32% |
| 21 | Literature-Letters | 19603 | 0.37% | 0.19% |
| 22 | Literature-Novels | 847945 | 16.20% | 8.37% |
| 23 | Literature-Plays | 86783 | 1.66% | 0.86% |
| 24 | Literature-Poetry | 9800 | 0.19% | 0.10% |
| 25 | Literature-Science Fiction | 63482 | 1.21% | 0.63% |
| 26 | Literature-Short Stories | 1242392 | 23.74% | 12.27% |
| 27 | Literature-Speeches | 83207 | 1.59% | 0.82% |
| 28 | Literature-Text Books (School) | 33804 | 0.65% | 0.33% |
| 29 | Literature-Travelogues | 159367 | 3.05% | 1.57% |
| 30 | Mythology | 17832 | 0.34% | 0.18% |
|  | **Total** | 5233452 | 100% | 51.68% |

**Table 10: Representation of Aesthetics**

## 11.2 Commerce

The Commerce text corpus covers 8 subdomains bearing a total of 66,924 words along with the overall percentage of 0.66%. The representational details are given in the table below:

| # | Sub domain | Word Count | Percentage (within Subdomain) | Overall Percentage |
|---|---|---|---|---|
| 1 | Accountancy | 5440 | 8.13% | 0.05% |
| 2 | Banking | 558 | 0.83% | 0.01% |
| 3 | Business | 7171 | 10.72% | 0.07% |
| 4 | Finance | 1499 | 2.24% | 0.01% |
| 5 | Industry | 11974 | 17.89% | 0.12% |
| 6 | Management | 35423 | 52.93% | 0.35% |
| 7 | Share Market | 1814 | 2.71% | 0.02% |
| 8 | Tourism | 3045 | 4.55% | 0.03% |
|  | **Total** | **66924** | **100%** | **0.66%** |

**Table 11: Representation of Commerce**

## 11.3 Mass Media

The Mass Media text corpus covers 18 sub-domains bearing a total of 33,54,996 words along with the overall percentage of 33.13%. The representational details are given in the table below:

| # | Sub domain | Word Count | Percentage (within Subdomain) | Overall Percentage |
|---|---|---|---|---|
| 1 | Article | 13508 | 0.40% | 0.13% |
| 2 | Business News | 3078 | 0.09% | 0.03% |
| 3 | Cinema News | 727 | 0.02% | 0.01% |
| 4 | Classifieds | 16099 | 0.48% | 0.16% |
| 5 | Discussions | 673602 | 20.08% | 6.65% |
| 6 | Editorial | 288547 | 8.60% | 2.85% |
| 7 | General News | 1756463 | 52.35% | 17.34% |
| 8 | Health | 5047 | 0.15% | 0.05% |
| 9 | Interviews | 42847 | 1.28% | 0.42% |
| 10 | Letters | 25164 | 0.75% | 0.25% |
| 11 | Obituary | 3848 | 0.11% | 0.04% |
| 12 | Political | 179385 | 5.35% | 1.77% |
| 13 | Religious/Spiritual News | 10230 | 0.30% | 0.10% |
| 14 | SMS | 1684 | 0.05% | 0.02% |
| 15 | Social | 70414 | 2.10% | 0.70% |
| 16 | Speeches | 88151 | 2.63% | 0.87% |
| 17 | Sports News | 174889 | 5.21% | 1.73% |
| 18 | Weather | 1313 | 0.04% | 0.01% |
|  | **Total** | 3354996 | 100% | 33.13% |

**Table 12: Representation of Mass Media**

## 11.4 Science and Technology

The Science and Technology text corpus covers 35 sub-domains bearing a total of 3,72,790 words along with the overall percentage of 3.68%. The quantitative representation is shown in the table below:

| # | Sub domain | Word Count | Percentage (within Subdomain) | Overall Percentage |
|---|---|---|---|---|
| 1 | Agriculture | 74524 | 19.99% | 0.74% |
| 2 | Astrology | 30746 | 8.25% | 0.30% |
| 3 | Astronomy | 8555 | 2.29% | 0.08% |
| 4 | Ayurveda | 39037 | 10.47% | 0.39% |
| 5 | Bio Chemistry | 1251 | 0.34% | 0.01% |
| 6 | Biology | 5353 | 1.44% | 0.05% |
| 7 | Biotechnology | 385 | 0.10% | 0.00% |
| 8 | Botany | 25290 | 6.78% | 0.25% |
| 9 | Chemistry | 4391 | 1.18% | 0.04% |
| 10 | Computer Sciences | 5264 | 1.41% | 0.05% |
| 11 | Criminology | 2248 | 0.60% | 0.02% |
| 12 | Engineering-Chemical | 1162 | 0.31% | 0.01% |
| 13 | Engineering-Civil | 502 | 0.13% | 0.00% |
| 14 | Engineering-Electrical | 2973 | 0.80% | 0.03% |
| 15 | Engineering-Electronics Communication | 1343 | 0.36% | 0.01% |
| 16 | Engineering-Others | 712 | 0.19% | 0.01% |
| 17 | Environmental Science | 12337 | 3.31% | 0.12% |
| 18 | Film Technology | 8985 | 2.41% | 0.09% |
| 19 | Forestry | 13533 | 3.63% | 0.13% |
| 20 | Geology | 9879 | 2.65% | 0.10% |
| 21 | Homeopathy | 2684 | 0.72% | 0.03% |
| 22 | Horticulture | 6256 | 1.68% | 0.06% |
| 23 | Logic | 2631 | 0.71% | 0.03% |
| 24 | Mathematics | 3092 | 0.83% | 0.03% |
| 25 | Medicine | 50832 | 13.64% | 0.50% |
| 26 | Micro Biology | 507 | 0.14% | 0.01% |
| 27 | Naturopathy | 1179 | 0.32% | 0.01% |
| 28 | Physics | 12590 | 3.38% | 0.12% |
| 29 | Psychology | 2438 | 0.65% | 0.02% |
| 30 | Sexology | 747 | 0.20% | 0.01% |
| 31 | Statistics | 2284 | 0.61% | 0.02% |
| 32 | Textile Technology | 10023 | 2.69% | 0.10% |
| 33 | Veterinary | 2567 | 0.69% | 0.03% |
| 34 | Yoga | 11184 | 3.00% | 0.11% |
| 35 | Zoology | 15306 | 4.11% | 0.15% |
|  | **Total** | 372790 | 100% | 3.68% |

**Table 13: Representation of Science and Technology**

## 11.5 Official Document

The Official Document text corpus covers 2 sub-domains bearing a total of 1,298 words along with the overall percentage of 0.01%. The representational details are given in the table below:

| Sub domain | Word Count | Percentage (within Subdomain) | Overall Percentage |
|---|---|---|---|
| Administration | 990 | 76.27% | 0.01% |
| Police Documents | 308 | 23.73% | 0.00% |
| Total | 1298 | 100% | 0.01% |

**Table 14: Representation of Official Document**

## 11.6 Social Science

The Social Science text corpus covers 22 sub-domains bearing a total of 10,97,570 words along with the overall percentage of 10.84%. The representational details are given in the table below:

| # | Sub domain | Word Count | Percentage (within Subdomain) | Overall Percentage |
|---|---|---|---|---|
| 1 | Anthropology | 2784 | 0.25% | 0.03% |
| 2 | Archaeology | 2860 | 0.26% | 0.03% |
| 3 | Demography | 3676 | 0.33% | 0.04% |
| 4 | Economics | 125527 | 11.44% | 1.24% |
| 5 | Education | 42643 | 3.89% | 0.42% |
| 6 | Food and Wellness | 19770 | 1.80% | 0.20% |
| 7 | Geography | 14683 | 1.34% | 0.14% |
| 8 | Health and Family Welfare | 77750 | 7.08% | 0.77% |
| 9 | History | 347753 | 31.68% | 3.43% |
| 10 | Home Science | 9568 | 0.87% | 0.09% |
| 11 | Journalism | 39815 | 3.63% | 0.39% |
| 12 | Law | 24695 | 2.25% | 0.24% |
| 13 | Linguistics | 44567 | 4.06% | 0.44% |
| 14 | Personality Development | 5566 | 0.51% | 0.05% |
| 15 | Philosophy | 59189 | 5.39% | 0.58% |
| 16 | Physical Education | 13677 | 1.25% | 0.14% |
| 17 | Political Science | 74949 | 6.83% | 0.74% |
| 18 | Public Administration | 6756 | 0.62% | 0.07% |
| 19 | Religion/Spiritual | 63025 | 5.74% | 0.62% |
| 20 | Sociology | 67308 | 6.13% | 0.66% |
| 21 | Sports | 44452 | 4.05% | 0.44% |
| 22 | Text Book (Social Science) | 6557 | 0.60% | 0.06% |
|  | **Total** | 1097570 | 100% | 10.84% |

**Table 15: Representation of Social Science**