EMILLE Corpus Documentation

This document contains:

This document contains a description of the structure of the corpus, as embodied in the filenames of the texts and the directory structure.

It also gives word counts for the various sections, and acknowledgements to the organisations and individuals who have allowed their texts to be used in the corpus

Encoding and markup

The text is encoded as two-byte Unicode text. For more information on Unicode, see www.unicode.org .

The texts are marked up in SGML using level 1 CES compliant markup. Each file also includes a full header, which specifies the provenance of the text. In the spoken corpus, information about the speakers is also stored in the header.

Corpus structure

There are three parts to the EMILLE Corpora. They are:

The EMILLE Spoken Corpus
The EMILLE-CIIL Monolingual Written Corpora
The EMILLE Parallel Corpus

However, for ease of use, the Spoken and Written Corpora are not organised separately in the directory structure. Instead, the spoken and written corpora for each language are grouped in a single directory, e.g. “hindi” for the Hindi spoken and written corpora, under the directory “monolingual”.

The parallel corpus is held separately, in the directory “parallel”.

Many of the directories contain subdirectories, in which texts are classified by a) their provenance or b) their genre. This structure is also implicit in the filename of each text within the corpora. While the directory structure and the filenames usually imply the same structure, there are some *ad hoc* exceptions to this, mentioned below where relevant.

A further section of the corpus contains *annotated* data, as discussed below.

Filenames

All files collected for the EMILLE Corpora have a filename in a standard format. Texts which have been incorporated from the CIIL Corpus have been given a filename in this standard format to render the two corpora compatible.

The filename consists of a series of codes chained together with hyphen characters. These codes specify the main language of the file, the source of the text, its subcategory in terms of subject matter if such information is available, and an identifying number. The name is generally of the format:

[Language]-[text type]-[datasource]-[subcategory]-[identifying number].txt

In the case of sources from which text was gathered on a periodical basis (i.e. the news websites in the written corpus, the radio programmes for the written corpus) the identifying number is a date. For other files it is simply an arbitrary distinguishing number.

For example:

hin-w-ranchi-news-01-03-22.txt
ben-s-cg-asiannet-02-07-23.txt

(see below for details on how language, text type and dates are signified.)

Files incorporated from the CIIL Corpus have a slightly different format. Since the CIIL Corpus data draws on a much wider variety of sources of data than the EMILLE Corpus, texts are sorted by their genre and subject matter, and uniquely identified by their original code in the CIIL Corpus, thus:

[Language]-[text type]-[genre]-[subcategory]-[CIIL Corpus code].txt

Some exceptions to this scheme are detailed in the discussion of the different parts of the corpora below. The major one is the Sinhala written corpus, which unlike the other languages is organised primarily by the category into which the text falls, and secondarily by the source it was gathered from.

Language codes

The codes used for languages in the filenames and also within the mark-up of some of the corpus files are drawn from ISO-639. See the table below.

Language	***Code***
Hindi	hin
Bengali	ben
Punjabi	pun
Gujarati	guj
Urdu	urd

Tamil	tam
Sinhala	sin
Marathi	mar
Oriya	ori
Assamese	asm
Kashmiri	kas
Malayalam	mal
Kannada	kan
Telugu	tel
English	eng

Text type

Text type	*Code*
Written	w
Spoken (demographically sampled)	s-dem
Spoken (context governed)	s-cg

Dates

All dates in filenames and anywhere in the header or markup of files in the corpora are given in the format **yy-mm-dd**.

The Annotated Data

Alongside the corpus is given some annotated data in Hindi and Urdu.

The data in Hindi consists of excerpts from the "Ranchi Express" data of the Hindi corpus (see below) annotated with anaphora analysis by Srija Sinha.

The data in Urdu consists of a copy of the Urdu written, spoken and parallel corpora annotated with morphosyntactic tags by the Urdu tagger created by Andrew Hardie.

The Parallel Corpus

The parallel corpus consists of 200,000 words of text in English and accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu.

The EMILLE Project would like to express our thanks to the UK Government and the various local authorities who generously gave us permission to incorporate their information leaflets into their parallel corpus.

For some texts, a version of some text was not available in one or more of the languages (either because the text was not translated in the first place, or because we were unable to locate a copy of the leaflet in the language in question). In this case, we have had translations made of the missing texts, either by employees at Lancaster or by an outside agency on our behalf. Where a translation is not "official" and has been produced by the EMILLE Project, this is indicated in the file header.

If two files are parallel to one another, their filenames are identical except for the language code at the start. So, *ben-w-housing-value* is parallel to *guj-w-housing-value*.

Filenames in the parallel corpus are more consistent that in the written corpus: each parallel set of files is assigned to a category (in the examples given above, "housing") and is given a unique identifier (like "value" above). This identifier is a word drawn from the title of the leaflet or summarising its contents.

The categories are:

consumer	Consumer issues
education	Education including school-home partnerships, special needs, etc.
housing	Government housing leaflets
health	NHS and public health information leaflets
legal	Legal issues including work permits, minimum wage etc.
social	Social Security and other social issues

Note that the categories are not evenly represented in the parallel corpus.

There follows a list of the unique identifiers, together with the titles and publishers of the leaflets they represent (this information is also contained in the header of the each relevant file).

Filename	Document Title	Publisher
attend	School attendance, information for parents	Department for Education and Employment
babies	Babies and Children BC1	Department of Social Security
bloodsample	If a blood sample is being taken...	Department of Health
breast	Be Breast Aware	Department of Health
buyers	A Buyer's / Shopper's Guide	Office of Fair Trading
cancer	Womens Nationwide Cancer Control Campaign	Womens Nationwide Cancer Control Campaign/Department of Health
catering	Assured Safe Catering	Department of Health
childcare	Childcare career	Department for Education and Skills
compensation	A better deal for tenants: Your new right to compensation for improvements	Department of the Environment, Transport and the Regions
consent	About the consent form	Department of Health
cot	Reducing The Risk Of Cot Death	Department of Health
county	Here to help you	Lancashire County Council

crime	Victims of crime	The Home Office
discharge	Discharge From Hospital	Manchester City Council
donor	Life Don't Keep It To Yourself	Department of Health
drugs	Drugs A Parent's Guide	Department of Health
drugschild	Drugs and Solvents: You and Your Child	Department of Health
drugsprobs	Services for people with drugs problems	Manchester City Council
exclusion	Preventing social exclusion (summary)	The Cabinet Office
eye	Eye Sight Tests	Department of Health
financial	Financial help if you work or are looking for work	Department of Social Security
foodlaw	Food Law Inspections	Ministry of Agriculture Fisheries and Food/Department of Health
haccp	Practical Food Safety for Businesses	Department of Health, Ministry of Agriculture, Fisheries and Food, and the Central Office of Information
headlice	The prevention and treatment of headlice	Department of Health
hepatitis	Hepatitis B	Department of Health
hiv	Services for people with HIV/AIDS	Manchester City Council
homeschool	Home-School agreements, What every parent should know (ISBN 0855229098)	Department for Education and Skills
landlord	My landlord wants me out	Department for Transport, Local Government and the Regions
law	Health and Safety Law	Health and Safety Executive
learning	Services for people with learning disabilities	Manchester City Council
littleread	A little reading goes a long way	Department for Education and Skills
liverpool	Unhappy with social services?	Liverpool Social Services
looking	How to get help in looking after someone	Department of Health
manage	A better deal for tenants: Your new right to manage	Department of the Environment, Transport and the Regions
manchester	How To Get Help From Social Services	Manchester City Council
markets	Modern markets: confident consumers	Department of Trade and Industry
maternity	Maternity services	Department of Health
meningitis	Knowing about Meningitis and Septicaemia	Department of Health
mmr	MMR The Facts	Department of Health
nation	The Health of the Nation and You	Department of Health
nhs	Help with NHS costs	Department of Health
noise	Bothered by noise	Department for Environment, Food and Rural Affairs
older	Health and Healthy Living: A guide for older people	Department of Health
ombudsman	The Health Service Ombudsman for England	Office of the Health Service Ombudsman
patients	The Patient's Charter and You	Department of Health
permit	Work Permits (UK) General Information	The Home Office
pregnant	While You Are Pregnant	Department of Health

race	New Laws... Race equality	The Home Office
readwrite	Learning to read and write at home and at school	Department for Education and Skills
rent	Do you rent, or are you thinking of renting, from a private landlord?	Department for Transport, Local Government and the Regions
repair	Your new right to repair	Department of the Environment
residential	Choosing Residential and Nursing Home Care	Manchester City Council
retire	Retirement RM1	Department of Social Security
rights	Your rights as a council tenant	Department of the Environment, Transport and the Regions
road	Teaching children road safety	Department of the Environment, Transport and the Regions
runaways	Consultation on Young Runaways (summary)	The Cabinet Office
scottish	Report of the Inquiry into the Liaison Arrangements between the Police, the Procurator Fiscal Service and the Crown Office and the Family of the Deceased Surjit Singh Chhokar in Connection with the Murder of Surjit Singh Chhokar and the Related Prosecutions (Dr Raj Jandoo).	Clerk of the Scottish Parliament/Scottish Parliament
senguide	Special Educational Needs (SEN) A guide for parents and carers	Department for Education and Skills
sentribunal	SEN tribunal: How to appeal	Department for Education and Skills
service	Work permits : Service and Standards	The Home Office
sick	Sick or disabled SD1	Department of Social Security
solvents	Solvents A parent's guide	Department of Health
supporting	Supporting People and Sheltered Housing	Department for Transport, Local Goverment and the Regions
teeth	Healthy teeth for life	Health Education Board for Scotland/National Health Service
tenant	Tenant participation compacts: a guide for tenants	Department of the Environment, Transport and the Regions
training	Work permits for training schemes and work experience	The Home Office
transport	Making the Connections: Transport and Social Exclusion (summary)	The Cabinet Office
tuberculosis	TB - are you aware?	Department of Health
value	Best Value in Housing	Department of the Environment, Transport and the Regions
vitamin	Vitamin K	Department of Health
wage	The National Minimum Wage Report Summary	Low Pay Commission
warm	Keep Warm Keep Well	Department of Health

Three texts are missing from the parallel corpus, which we were not able to get transcribed. They are: *liverpool* (Bengali), *manchester* (Hindi), and *pregnant* (Bengali).

The parallel corpus contains full sentence markup using the <s> element, which is not the case for the majority of files within the written corpus.

The Spoken Corpus

As mentioned above, for ease of use the different parts of the Spoken Corpus are grouped with the same-language written texts

Most of the data in these corpora is context-governed speech (transcripts of radio programmes from the BBC Asian Network). The Bengali and Hindi corpora also contain small amounts of demographically-sampled speech. This is indicated in the filename, as specified above (e.g. *ben-**s-dem**-302.txt* as opposed to *ben-**s-cg**-asiannet-02-11-26.txt).*

The context-governed files are further subdivided according to the radio programme from which they were derived. This is the third element of the filename. For Gujarati and Punjabi, there is no subdivision as all texts were derived from the BBC Asian Network Gujarati and Punjabi programmes.

The date of the original broadcast completes the filename. The names of the programmes as abbreviated in the filenames are listed and expanded upon below:

asiannet	The BBC Asian Network Language Programmes (all languages: broadcast nightly between 7.30 and 10 p.m.)
afternoon	The Afternoon Show with Navinder Bhogal (Hindi and Urdu: weekday afternoons 2-4 p.m.
shamoly	The BBC Radio Lancashire *Shamoly* programme (Bengali: Sunday evenings)
jaltrang	The BBC Radio Lancashire *Jaltrang* programme (Urdu: Sunday evenings)

The size of the Spoken Corpus is as follows:

Language	Word count
Bengali	442,000
Hindi	588,000
Urdu	512,000
Gujarati	564,000
Punjabi	521,000
Total	2,628,000

The EMILLE-CIIL Monolingual Written Corpora

Filenames in the written corpus consist of codes for the language and text type, as specified above, followed by a word to specify the source of the data, possibly followed by a subcategory. At the end of the filename is the date of publication of the original text (in the case of news data) or a code number (for other files).

The written corpus incorporates the CIIL Corpus, created by the **Central Institute of Indian Languages, Mysore** in collaboration with the **Indian Institute of Technology, Delhi**, the **Institute of Applied Language Sciences, Bhubaneshwar**, and **Aligarh Muslim University, Aligarh**.

This corpus, originally encoded as ISCII text, has been re-encoded as Unicode with CES-compliant SGML markup, as per the data collected by the EMILLE Project, to allow simultaneous use of both datasets.

The filenames of texts from the CIIL Corpus differ slightly from the overall scheme. Because the CIIL Corpus was drawn from a much wider set of genres, and from many, many more sources than the EMILLE Project's data, it was not considered suitable to classify them by source. Instead they are classified by genre (category and subcategory). The filename is concluded by the code identifying that file in the original CIIL Corpus. Within the corpus structure these files are grouped in a directory entitled "miscellaneous" (because the data derives from a wide miscellany of sources).

The EMILLE-CIIL Monolingual Written Corpora have a total size of approximately 93,530,000 words. The make-up of each of the fourteen language corpora is discussed below.

Hindi

The EMILLE Project would like to express our thanks to the text providers who generously allowed us to gather data from their websites:

- IndiaInfo
- Ranchi Express
- Webdunia

The Hindi corpus also contains data incorporated from the CIIL Corpus, originally gathered by the Indian Institute of Technology.

The contents of the Hindi written corpus are as follows (broken down by directory):

- Data from the **IndiaInfo** news website: approximately 600,000 words
- Data from the **Webdunia** news website: approximately 5,600,000 words

- Data from the **Ranchi Express** news website: approximately 3,200,000 words
- Data from miscellaneous sources (incorporated from the CIIL Corpus): approximately 3,000,000 words.

The Hindi written corpus contains a total of approximately 12,390,000 words.

Bengali

The EMILLE Project would like to express our thanks to the text providers who generously allowed us to gather data from their websites:

- Bengalnet

The Bengali corpus also contains data incorporated from the CIIL Corpus, originally gathered by the Institute of Applied Language Sciences, Bhubaneshwar.

The contents of the Bengali written corpus are as follows (broken down by directory):

- Data from the **Bengalnet** news website : approximately 1,980,000 words
- Data from miscellaneous sources (incorporated from the CIIL Corpus): approximately 3,540,000 words.

The Bengali written corpus contains a total of approximately 5,520,000 words.

Punjabi

The EMILLE Project would like to express our thanks to the text providers who generously allowed us to gather data from their websites:

- Daily Ajit, Jalandhar
- Sanjh Savera
- Eh Din
- Nagara

We would also like to express our thanks to the staff of the *Panchim* periodical, who supplied us with copies of their text, and to the "Gurbani CD" project, who generously allowed us to use their CD as the source for the text of the Shree Guru Granth Sahib.

The Punjabi corpus also contains data incorporated from the CIIL Corpus, originally gathered by the Indian Institute of Technology.

Note that, unlike the other languages, the Punjabi written corpus contains data in more than one writing system. The data from the *Panchim* periodical is written in the Indo-Perso-Arabic ("Shahmukhi") alphabet, whereas the remainder of the data is written in the Gurmukhi alphabet. The two different types of written data are held in separate directories.

The contents of the Punjabi written corpus are as follows (broken down by directory):

- Data from the **Panchim** periodical (in Indo-Perso-Arabic script): approximately 3,130,000 words.
- Full text of the **Shree Guru Granth Sahib** in a single file: approximately 510,000 words
- Data from the **Eh Din** website: approximately 90,000 words
- Data from the **Nagara** website: approximately 100,000 words
- Data from the **Sanjh Savera** website: approximately 1,200,000 words**
- Data from the **Daily Ajit** website: approximately 8,500,000 words
- Data from **miscellaneous** sources (incorporated from the CIIL Corpus): approximately 1,970,000 words.

** Note that due to difficulties in the encoding of the text, it was not possible to obtain an accurate word-count for the data drawn from the Sanjh Savera website. The figure above represents an educated estimate based on a sample of the data.

The Punjabi written corpus contains a total of approximately 15,600,000 words.

Gujarati

The EMILLE Project would like to express our thanks to the text providers who generously allowed us to gather data from their websites:

- Gujarat Samachar.

The Gujarati corpus also contains data incorporated from the CIIL Corpus, originally gathered by the Central Institute for Indian Languages.

The contents of the Gujarati written corpus are as follows (broken down by directory):

- Data from the **Gujarat Samachar** website, divided into parts:
    - **general** – data from various parts of the site (approximately 390,000 words)
    - **national** – national news from India (approximately 750,000 words)
    - **news** – news data of mixed provenance (approximately 690,000 words)
    - **regional** – state news from Gujarat (approximately 3,930,000 words)
    - **supplement** – data collected from the daily "supplement" pages of the site, containing a mixture of features and other articles (approximately 4,740,000 words)
- Data from **miscellaneous** sources (incorporated from the CIIL Corpus): approximately 1,500,000** words.

(** This is a very imprecise word-count.)

The Gujarati written corpus contains a total of approximately 12,150,000 words.

Urdu

The Urdu written corpus consists of data incorporated from the CIIL Corpus, originally gathered by Aligarh Muslim University (approximately 1,640,000 words).

Tamil

The EMILLE Project would like to express our thanks to the text providers who generously allowed us to gather data from their websites:

- Dinakaran

The Tamil corpus also contains data incorporated from the CIIL Corpus, originally gathered by the Central Institute for Indian Languages.

The contents of the Tamil written corpus are as follows (broken down by directory):

- Data from the **Dinakaran** website, divided into parts:
  - **cinema** – film and related news (approximately 1,050,000 words)
  - **news** – news stories (approximately 8,610,000 words)
  - **other** – other stories from the website (approximately 730,000 words)
  - **politics** – political news and commentary (approximately 5,050,000 words)
  - **sports** – sports news reportage (approximately 1,170,000 words)
- Data from miscellaneous sources (incorporated from the CIIL Corpus): approximately 3,380,000 words.

The Tamil written corpus contains a total of approximately 19,980,000 words.

Sinhala

The Sinhala corpus consists of data collected from the press and media of Sri Lanka by Vincent Halahakone. The EMILLE Project would like to express our thanks to the text providers who generously allowed their data to be incorporated into the corpus and to all the other parties and individuals who helped with the Sinhala text collection, including:

- The Associated Newspapers of Ceylon Ltd, Colombo, Sri Lanka
- Upali Newspapers Ltd, Colombo, Sri Lanka
- Wijesooriya Book Centre, Colombo, Sri Lanka
- Sara Publishers, Colombo, Sri Lanka
- Ravaya Publishers (Guarantee) Ltd. Colombo, Sri Lanka
- Movemnet for the Defence of Democracy, Colombo, Sri Lanka
- Diyasa Study Circle, Colombo, Sri Lanka
- Department of Cultural Affairs, Colombo, Sri Lanka

- Dr. A. Obeysekara, Colombo, Sri Lanka
- Prof Gamini Adhikari , Colombo, Sri Lanka
- Post Graduates Institute of Archaeology, Colombo, Sri Lanka
- Yukthiya Newspapers Ltd., Colombo, Sri Lanka
- Centre for Participatory Development, Colombo, Sri Lanka
- Pirivena Education Division, Ministry of Education and Higher Education. Colombo, Sri Lanka
- Central Cultural Fund, Colombo Sri Lanka
- A.B. Kotalawala, University of Sri Jayawardenapura, Colombo Sri Lanka
- Sri Lanka National Film Corporation, Colombo, Sri Lanka
- Dr. Daya Rohana Atukorala, University of Colombo
- Independent Financial News & Views (PVT) Ltd, Colombo, Sri Lanka
- Sarvodaya Vishvalekha, Sarvodaya, Sri Lanka
- Divaina internet news

The Sinhala corpus is structured slightly differently to the other written corpora (and this is reflected in the filenames and directory structure). The primary classification is by text type (see below). The source is the secondary classification. Subsequent elements in the filename may indicate a further subcategory or an element of the title. Filenames end either with a code number or a date, as per usual.

The contents of the Sinhala written corpus are as follows (broken down by directory):

- "literature" – fiction (novels and short stories) and biography: approximately 1,200,000 words.
- "editorials" – editorial material from the Sinhala media: approximately 100,000 words
- "features" – feature articles on a wide range of topics: approximately 1,400,000 words
- "humanities" – humanities articles: approximately 600,000 words
- "news" – subdivided into "home" and "foreign": approximately 1,800,000 words
- "webnews" – news data gathered from internet sites: approximately 1,760,000 words

The Sinhala written corpus contains a total of approximately 6,860,000 words.

Marathi

The Marathi written corpus consists of data incorporated from the CIIL Corpus, originally gathered by the Central Institute for Indian Languages (approximately 2,210,000 words).

Oriya

The Oriya written corpus consists of data incorporated from the CIIL Corpus, originally gathered by the Institute of Applied Language Sciences, Bhubaneshwar (approximately 2,730,000 words).

Assamese

The Assamese written corpus consists of data incorporated from the CIIL Corpus, originally gathered by the Institute of Applied Language Sciences, Bhubaneshwar (approximately 2,620,000 words).

Kashmiri

The Kashmiri written corpus consists of data incorporated from the CIIL Corpus, originally gathered by Aligarh Muslim University (approximately 2,270,000 words).

Malayalam

The Malayalam written corpus consists of data incorporated from the CIIL Corpus, originally gathered by the Central Institute for Indian Languages (approximately 2,350,000 words).

Kannada

The Kannada written corpus consists of data incorporated from the CIIL Corpus, originally gathered by the Central Institute for Indian Languages (approximately 2,240,000 words).

Telugu

The Telugu written corpus consists of data incorporated from the CIIL Corpus, originally gathered by the Central Institute for Indian Languages (approximately 3,970,000 words).