Corpus Data for South Asian Language Processing

Paul Baker, Andrew Hardie, Tony McEnery
Department of Linguistics, Lancaster University, UK
{j.p.baker, a.hardie, a.mcenery}@lancaster.ac.uk

B.D. Jayaram
Central Institute of Indian Languages, Mysore, India
jayaram@ciil.stpmy.soft.net

Abstract

The EMILLE Project (Enabling Minority Language Engineering) was established to construct a 67 million word corpus of South Asian languages. In addition, the project has had to address a number of issues related to establishing a language engineering (LE) environment for South Asian language processing, such as translating 8-bit language data into Unicode and producing a number of basic LE tools. This paper will focus on the corpus construction undertaken on the project and will outline the rationale behind data collection. In doing so a number of issues for South Asian corpus building will be highlighted.

1 Introduction

The EMILLE project\(^1\) has three main goals: to build corpora of South Asian languages, to extend the GATE LE architecture\(^2\) and to develop basic LE tools. The architecture, tools and corpora should be of particular importance to the development of translation systems and translation tools. These systems and tools will, in turn, be of direct use to translators dealing with languages such as Bengali, Hindi and Punjabi both in the UK and internationally (McEnery, Baker and Burnard, 2000).

This paper discusses progress made towards the first of these goals and considers to a lesser extent the third goal of the project. Readers interested in the second goal of the project are referred to Tablan et al (2002).

2 Development of the corpora

This section describes our progress in collecting and annotating the different types of corpora covered by EMILLE. EMILLE was established with the goal of developing written language corpora of at least 9,000,000 words for Bengali, Gujarati, Hindi, Punjabi, Sinhalese, Tamil and Urdu. In addition, for those languages with a UK community large enough to sustain spoken corpus collection (Bengali, Gujarati, Hindi, Punjabi and Urdu), the project aimed to produce spoken corpora of at least 500,000 words per language and 200,000 words of parallel corpus data for each language based on translations from English. At the outset we decided to produce our data in Unicode and annotate the data according to the Corpus Encoding Standard (CES) guidelines\(^3\). As the project has developed, the initial goals of EMILLE have been refined. In the following subsections we describe the current

\(^1\) Funded by the UK EPSRC, project reference GR/N19106. The project commenced in July 2000 and is due to end in September 2003.

\(^2\) Funded by the UK EPSRC, project references GR/K25267 and GR/M31699.

\(^3\) Readers interested in the markup of the corpus are referred to Baker et al (2002).
state of the EMILLE corpora and outline the motives behind the various refinements that have been made to EMILLE’s goals.

2.1 Monolingual written corpora

The first major challenge facing any corpus builder is the identification of suitable sources of corpus data. Design criteria for large scale written corpora are of little use if no repositories of electronic text can be found with which to economically construct the corpus. This causes problems in Indic corpus building as the availability of electronic texts for Indic languages is limited. This availability does vary by language, but even at its best it cannot compare with the availability of electronic texts in English or other major European languages. We realised that much of the data which, in principle, we would have liked to include in the corpus existed in paper form only. On EMILLE, it would have been too expensive to pay typists to produce electronic versions of the 63 million words of monolingual written corpus (MWC) data. Even if the initial typing had been affordable, checking the data for errors would have added a further cost, particularly since tools for error correction, such as spell checkers, do not exist for many of the languages studied on EMILLE (Somers, 1998, McEnery and Ostler, 2000). Scanning in the text using an optical character recognition (OCR) program is a viable alternative to typing in printed text for languages printed in the Roman alphabet. However, OCR programs for Indic scripts are still in their infancy (for an example of some early work see Pal and Chaudhuri, 1995) and were not considered stable and robust enough for this project to use gainfully.4

As part of a pilot project to EMILLE5, we ran a workshop that examined potential sources of electronic data for Indian languages. The workshop identified the Internet as one of the most likely sources6. This prediction proved accurate, and we have gathered our MWC corpus from the web on the basis of four, largely pragmatic, criteria:

1. Data should only be gathered from sources which agreed to the public distribution of the data gathered for research purposes;
2. Text must be machine readable: we could not afford to manually input tens of millions of words of corpus data;
3. Each web-site used should be able to yield significant quantities of data: to focus our efforts we excluded small and/or infrequently updated websites from our collection effort;
4. Text should be gathered in as few encoding formats as possible: as we map all data to Unicode, we wished to limit the amount of mapping software we needed to author to achieve this task.

While the first three criteria are somewhat easy to understand and have been discussed elsewhere (Baker et al, 2002) the fourth criteria merits some discussion. Ideally, we would have liked to include texts that already existed in Unicode format in our corpus. However, when we first started to collect data, we were unable to locate Indic documents that had been created in Unicode.7 We found that creators of Indic documents on the internet typically rely on five methods for publishing texts online:

They use online images, usually in GIF or JPEG format. Such texts would need to be keyed in again, making the data of no more use to us than a paper version;

- They publish the text as a PDF file. Again, this made it almost impossible to acquire the original text in electronic format. We were sometimes able to acquire ASCII text from these documents, but were not able to access the fonts that had been used to create the Indic script texts. Additionally, the formatting meant that words in texts would often appear in a jumbled order, especially when acquired from PDF

---

4 We wished to produce the corpora in the original scripts and hence avoided Romanised texts altogether.

5 This project, Minority Language Engineering (MILLE), was funded by the UK EPSRC (Grant number GR/L96400).

6 While we also considered publishers of books, religious texts, newspapers and magazines as a possible data source, the prevalence of old-fashioned hot-metal printing on the subcontinent made us realise early on that such sources were not likely providers of electronic data. Indeed, a number of publishers expressed an interest in helping us, but none could provide electronic versions of their texts.

7 To date, the only site we have found that uses Unicode for Indic languages is the BBC’s; see for example www.bbc.co.uk/urdu or www.bbc.co.uk/hindi.
documents that contained tables, graphics or two or more columns;
• They use a specific piece of software in conjunction with a web browser. This was most common with Urdu texts, where a separate program, such as Urdu 98, is often used to handle the display of right-to-left text and the complex rendering of the nasta’liq style of Perso-Arabic script;
• They use a single downloadable True Type (TTF) 8-bit font. While the text would still need to be converted into Unicode, this form of text was easily collected;
• They use an embedded font. For reasons of security and user-convenience, some site-developers have started to use OpenType (eot) or TrueDoc (pfr) font technology with their web pages. As with PDF documents, these fonts no longer require users to download a font and save it to his or her PC. However, gaining access to the font is still necessary for conversion to Unicode. Yet gathering such fonts is difficult as they are often protected. We found that owners of websites that used embedded fonts were typically unwilling to give those fonts up. Consequently using data from such sites proved to be virtually impossible.

There are a number of possible reasons for the bewildering variety of formats and fonts needed to view Indic language data on the web. For example, many news companies who publish Indic language data on the web use in-house fonts or other unique rendering systems, possibly to protect their data from being used elsewhere, or sometimes to provide additional characters or logos that are not part of ISCII. However, the obvious explanation for the lack of Unicode data is that, to date, there have been few Unicode-compliant word-processors available. Similarly, until the advent of Windows 2000, operating systems capable of rendering Indic Unicode data successfully were not in widespread use. Even where a producer of data had access to a Unicode word-processing/web-authoring system they would have been unwise to use it, as the readers on the web were unlikely to be using a web browser which could successfully read Unicode and render Indic scripts.

Given the complexities of collecting this data, we chose to collect text from Indian language websites that offered a single downloadable 8-bit TTF font. Unlike fonts that encode English, such as Times New Roman as opposed to Courier, Indic fonts are not merely repositories of a particular style of character rendering. They represent a range of incompatible glyph encodings. In different English fonts, the hexadecimal code 0042 is always used to represent the character “B”. However, in various fonts which allow one to write in Devanagari script (used for Hindi among other languages), the hexadecimal code 0042 could represent a number of possible characters and/or glyphs. While ISCII (Bureau of Indian Standards, 1991) has tried to impose a level of standardisation on 8-bit electronic encodings of Indic writing systems, almost all of the TTF 8-bit fonts have incompatible Indian glyph encodings (McEnery and Ostler, 2000). ISCII is ignored by Indic TTF font developers and is hence largely absent from the web. To complicate matters further, the various 8-bit encodings of Indic writing systems have different ways of rendering diacritics, conjunct and half-form characters. For example, the Hindi font used for the online newspaper Ranchi Express tends to only encode half-forms of Devanagari, and a full character is created by combining two of these forms together. For example, to produce he (U+092A – Devanagari character he) in this font, two keystrokes would need to be entered (h + e).However, other fonts may use a single keystroke to produce he.

We were mindful that for every additional source of data using a new encoding that we wished to include in our corpus, an additional conversion table would have to be written in order to convert that corpus data to the Unicode standard. This issue, combined with the scarcity of existing Indic electronic texts, meant that we didn’t use as many sources of data as we would have initially liked, meaning we had to focus almost exclusively on newspaper material. However, as is noted in the following paragraph, as a consequence of the collaboration between Lancaster University and the Central Institute of
Indian Languages (CIIL), the eventual corpus will now contain a wider range of genres. Web data gathered on the basis of these four criteria would have allowed us to fulfil our original MWC project goals. However, the MWC collection goals of the project have altered significantly. Thanks to a series of grants from the UK EPSRC\(^8\) the project has been able to establish a dialogue with a number of centres of corpus building and language engineering research in South Asia. As a consequence, the EMILLE team has joined CIIL in Mysore, India, with the goal of producing a wider range of monolingual written corpora than originally envisaged on the EMILLE project. One effect of this change means that the uniform word counts of the monolingual written corpora will be lost.\(^9\) Each language will now be provided with varying amounts of data, though no language will be furnished with less than two million words. However, there is a further important effect of this collaboration: the corpus will now be able to cover a much wider range of languages (14 rather than 7) and a wider range of genres. By a process of serendipity, the corpus data being provided by CIIL covers a wide range of genres but not newspaper material.\(^10\) As the material gathered at Lancaster focuses exclusively on newspapers, the CIIL and Lancaster data is complementary. Table 1 shows the state of the EMILLE/CIIL monolingual written corpora at present, and the revised target corpus size.

The collection phase for the EMILLE/CIIL MWC data is nearly finished. Only around 13 million words of data remain to be collected.\(^11\) Consequently, the focus of the project is now falling increasingly on parallel and spoken data.

### Table 1: Word counts for each language in the EMILLE/CIIL Corpus as of January 2003

<table>
<thead>
<tr>
<th>Language</th>
<th>Target word count (millions)</th>
<th>Current word count (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assamese</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Bengali</td>
<td>9.0</td>
<td>5.5</td>
</tr>
<tr>
<td>Gujarati</td>
<td>10.6</td>
<td>10.6</td>
</tr>
<tr>
<td>Hindi</td>
<td>12.0</td>
<td>11.2</td>
</tr>
<tr>
<td>Kannada</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>Kashmiri</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Malayalam</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Marathi</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>Oriya</td>
<td>2.7</td>
<td>2.7</td>
</tr>
<tr>
<td>Punjabi</td>
<td>9.0</td>
<td>4.5</td>
</tr>
<tr>
<td>Sinhalese</td>
<td>9.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Tamil</td>
<td>15.0</td>
<td>13.9</td>
</tr>
<tr>
<td>Telegu</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Urdu</td>
<td>3.0</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>85.9</strong></td>
<td><strong>72.1</strong></td>
</tr>
</tbody>
</table>

### 2.2 Parallel corpora

The problems we encountered in collecting MWC data were also encountered when we started to collect parallel data. However, the relatively modest size of the parallel corpus we wished to collect (200,000 words in six languages) meant that we were able to contemplate paying typists to produce electronic versions of printed parallel texts. We eventually decided to do this as we had an excellent source of parallel texts which covered all of the languages we wished to look at: UK government advice leaflets. This was a good source of data for us, as we wished to collect data relevant to the translation of Indic languages in UK in a genre that was term rich.

The leaflets we were able to gather were mostly in PDF or print-only format, though some also used 8-bit encodings. Typing these texts became a necessity when the UK government gave us permission to use the texts, but the company that produced the electronic versions of the texts refused to give us the electronic originals. We found it economic to pay typists to produce Unicode versions of the texts using Global Writer, a Unicode word-processor.\(^12\)

---

\(^8\) Grants GR/M70735, GR/N28542 and GR/R42429/01.

\(^9\) This change was also necessitated by the varying availability of suitable newspaper websites for the different languages. For Hindi and Tamil, for example, plenty of data is available to be gathered; for Punjabi and Bengali, somewhat less; for Urdu, almost none.

\(^10\) The data provided by CIIL to the project covers a number of genres, including Ayurvedic medicine, novels and scientific writing.

\(^11\) 3.5 million words of Bengali, 2.1 of Hindi, 5.2 of Punjabi, 2.8 of Tamil, 4 of Sinhalese and 1.4 of Urdu.

\(^12\) When the project began, Global Writer was one of the few word-processors able to handle the rendering of Indic languages in Unicode. Since then, Microsoft have made Word 2000 Unicode-compliant. However, unless...
The research value of the British government data is very high in our view. The UK government produces a large number of documents in a wide range of languages. All are focused in areas which are term-rich, e.g. personal/public health, social security and housing. To build the parallel corpus we collected about 75 documents from the Departments of Health, Social Services, Education and Skills, and Transport, Local Government and the Regions.\footnote{We were also able to take a smaller number of texts from the Home Office, the Scottish Parliament, the Office of Fair Trading, and various local government bodies (e.g. Manchester City Council).}

Other than the need to type the data from paper copies, the parallel corpus also presented one other significant challenge: while most of the data is translated into all of the languages we need, there are a few instances of a document not being available in one of the languages. Our solution is to employ translators to produce versions of the documents in the appropriate language. While far from ideal, this is not unprecedented as the English Norwegian Parallel Corpus project also commissioned translations (see Oksefjell, 1999). All such texts are identified as non-official translations in their header.

The parallel corpus is now complete, and we are beginning the process of sentence aligning the texts using the algorithm of Piao (2000).

\subsection*{2.3 Spoken corpora}

For the collection of spoken data we have pursued two strategies. Firstly we explored the possibility of following the BNC (British National Corpus) model of spoken corpus collection (see Crowdy, 1995). We piloted this approach by inviting members of South Asian minority communities in the UK to record their everyday conversations. In spite of the generous assistance of radio stations broadcasting to the South Asian community in the UK, notably BBC Radio Lancashire and the BBC Asian Network, the uptake on our offer was dismal. One local religious group taped some meetings conducted in Gujarati for us, and a small number of the people involved in typing work on the project agreed to record conversations with their family and friends. The feedback from this trial was decisive – members of the South Asian minority communities in Britain were uneasy with having their everyday conversations included in a corpus, even when the data was fully anonymised. The trial ended with only 50,000 words of spoken Bengali and 40,000 words of Hindi collected in this way.

Consequently we pursued our second strategy and decided to focus on Asian radio programmes broadcast in the UK on the BBC Asian Network as our main source of spoken data.\footnote{Programmes broadcast in Bengali and Urdu on BBC Radio Lancashire make up the remainder of the spoken corpus.} The BBC Asian Network readily agreed to allow us to record their programmes and use them in our corpus. The five languages of the EMILLE spoken corpora (Bengali, Gujarati, Hindi-Urdu, and Punjabi) are all covered by programmes on the BBC Asian Network. At least four and a half hours in each language (and more in the case of Hindi-Urdu) are broadcast weekly. The programmes play Indian music (the lyrics of which have not been transcribed) as well as featuring news, reviews, interviews and phone-ins. As such the data allows a range of speakers to be represented in the corpus, and some minimal encoding of demographic features for speakers is often possible as at least the sex of the speaker on the programmes is apparent.

The recordings of the radio programmes are currently being digitised and edited, to remove songs and other such material. The recordings will be made available in conjunction with the transcriptions. However, the transcriptions and recordings will not be time aligned. An obvious future enhancement of this corpus data would be to work on techniques, already well established for English, to time align the transcriptions.

The recording and transcription of the broadcasts is ongoing and to date we have completed the transcription of 265,000 words of Bengali, 109,000 words of Gujarati, 41,000 words of Hindi, 119,000 words of Urdu.

\section*{4 Corpus Annotation}

The corpus annotation research of EMILLE has recently expanded to cover another form of annotation – the annotation of demonstratives – in Hindi. The work on Hindi is at an early stage.
with an annotation scheme originally designed to annotate demonstratives in English (Botley & McEnery, 2001) being used to annotate Hindi. The annotation is currently underway and the goal is to annotate the demonstratives in 100,000 words of Hindi news material by the end of the project. However, the project always intended from the outset to explore morphosyntactic corpus annotation of Urdu. The work undertaken on this is covered in the next section.

4.1 Morphosyntactic Annotation

On the EMILLE project we wished to develop a POS tagger for at least one of the languages covered by the project. The language we have chosen to focus on is Urdu. We selected Urdu for a number of reasons. Firstly, it is widely spoken in the UK, both as a first and second language, and native speakers were available to be consulted at Lancaster where this part of the project is taking place. Secondly, as the lingua franca of a multilingual community (that of South Asian Muslims) and the official language of Pakistan, Urdu has considerable political and cultural importance. Thirdly, there are a number of factors that we anticipated would make tagging Urdu more complicated than tagging any other EMILLE language. For example, the right-to-left directionality of the Perso-Arabic script in which Urdu is written and the presence of grammatical forms borrowed from Arabic and Persian, which are structurally quite distinct from Indo-Aryan forms, mean that Urdu represents a unique challenge in our data. It seemed the best course of action to confront these problems by choosing Urdu as the language for which to develop POS tagging.

The first stage of the work was to develop a tagset for use in Urdu texts and corpora. The next stage, now underway, is to test the tagset’s usability in manual tagging, and build up a set of tagged texts to serve as training data for the final phase of this part of the project. This will be to automate the tagging and subsequently tag the whole of the Urdu corpus. In this section, we discuss the first, completed stage of this process, in which a tagset for Urdu was devised using the Urdu grammar of Schmidt (1999) as a basis.

The tagset was created in accordance with the EAGLES guidelines on morphosyntactic annotation (Leech and Wilson, 1999). These guidelines were designed to help standardise tagsets for the official languages of the European Union. While Urdu did not fall under the EAGLES remit, it was decided to work with this international standard in order to ensure the maximum utility of the final tagged corpus. Also, from a typological perspective it is not unreasonable to expect that the EAGLES guidelines would prove compatible with Urdu on the grounds that both Urdu and the original EAGLES languages were all of the Indo-European family. Indeed, it transpired that most of categories in the attribute-value system outlined in the EAGLES guidelines were suitable for application in the design of the Urdu tagset. There was no major group of Urdu words for which there was no equivalent category in EAGLES. The EAGLES guidelines deal very well with the gender, case and number system of Urdu and need only minor modifications – for example, since there was no value for oblique case in the EAGLES system, the value for dative case was used instead, on the grounds that the usage of the Urdu oblique corresponds quite closely to that of the dative in some EU languages, such as German. The verbal system proved a little more problematic, in the sense that the mood, tense and finiteness features outlined in the EAGLES attribute-value system do not map easily onto those found in the Urdu language. However, the greatest difficulty arose in dealing with the minor, idiosyncratic features of Urdu – whilst the idiosyncratic features of the EU languages are covered by the EAGLES guidelines this is not the case for Urdu. These features include: the appearance of case on some verbal elements; the distinction between ‘marked’ and ‘unmarked’ nouns; the Urdu honorific pronoun, which does not fit easily into any of the EAGLES categories for pronouns;

---

15 Urdu has masculine and feminine gender, singular and plural number, and nominative and oblique case, all expressed in a single fusional suffix on each noun/adjective.
16 Urdu verbs have one simple finite verb form (the subjunctive), two simple forms that may be finite or non-finite (the perfective and imperfective participles), and two further non-finite simple forms (the root and the infinitive). There are, however, a large number of complex verb forms using auxiliary elements.
17 The participles and the infinitive can all display case.
the borrowed Persian enclitic called ʾizāka; and the problem of bound derivational suffixes which appear in some contexts as independent tokens, but not in others. However, none of these problems were insurmountable. EAGLES has proved a robust and useful framework within which to approach Urdu POS tagging.

5 Accessing the corpus

A beta release of the EMILLE/CIIL corpus will be available, free of charge, for users from April 2003. The beta release of the corpus will contain a sample of MWC, parallel and spoken data for the core EMILLE language. In order to register for access to the beta release, users should contact Andrew Hardie.

6 Conclusion

The EMILLE project has adapted and changed over the course of the past two years. With regard to the EMILLE corpora, this has in large part been due to the project team engaging in a dialogue with the growing community of researchers working on South Asian languages. As a result of this dialogue the EMILLE team has made some major changes to the original design of the EMILLE corpora. However, as with all large-scale corpus-building projects, other changes have occurred on the project which have been responses to unexpected factors, such as the reluctance of members of the minority communities to engage in the recording of everyday spontaneous speech, and the lack of compatible 8-bit font encoding standards used by the different producers of Indic electronic texts. Devising methodologies to convert the numerous disparate 8-bit based texts to Unicode has been one of the most complex and time-consuming tasks of the project.

The area of South Asian corpus building is growing. As well as work in the UK and India, a new centre for South Asian language resources has been established in the US. As the centres cooperate and integrate their research, there is little doubt that further work on the construction and annotation of South Asian corpora will grow. As this work grows, I believe that corpus builders should not lose sight of two important truths. Firstly, that collaboration is better than competition – the corpus produced by Lancaster/CIIL will be larger and better because we have accepted this. The construction of large scale language resources needs the acceptance of this truth if it is to be effective. Secondly, that while many South Asian languages are entering the growing family of languages for which corpus data is available, there are still languages spoken in South Asia and the world for which corpus data is not available. While we must celebrate the creation of corpora of Indic languages, we should also think of the work yet to be done in creating corpora for those languages not yet corpus enabled.

References


