# EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation

**Paul Baker*, Andrew Hardie*, Tony McEnery*, Hamish Cunningham†, Rob Gaizauskas†**

* Dept. Linguistics, Lancaster University, Lancaster, UK
{j.p.baker, a.hardie, a.mcenery}@ilsp.gr

†Dept. Computer Science, Sheffield University, Sheffield, UK
{hamish, r.gaizauskas} @dcs.shef.ac.uk

**Abstract**

The paper describes developments to date on the EMILLE Project (Enabling Minority Language Engineering) being carried out at the Universities of Lancaster and Sheffield. EMILLE was established to construct a 67 million word corpus of South Asian languages. In addition to undertaking this corpus construction, the project has had to address a number of related issues in the context of 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. The development of tools on EMILLE has contributed to the on-going development of the LE architecture GATE.

## 1. Introduction

Our project has three main goals: to build corpora of South Asian languages, to extend the GATE[1] LE architecture and to develop basic LE tools. These three goals, when met, 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 Bangla, Hindi and Panjabi both in the UK and internationally (McEnery, Baker & Burnard, 2000). The project commenced in July 2000 and is due to end in September 2003. Below we report on progress on EMILLE to date.

## 2. Development of the corpora

This section describes our progress in collecting and annotating the different types of corpora covered by EMILLE. EMILLE[2] was established with the goal of developing written language corpora of at least 9,000,000 words for Bangla, Gujarati, Hindi, Panjabi, Singhalese, Tamil and Urdu. In addition, for those languages with a UK community large enough to sustain spoken corpus collection (Bangla, Gujarati, Hindi, Panjabi 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, wherever possible, in Unicode and annotate the data according to the Corpus Encoding Standard (CES) guidelines. As the project has developed, the initial goals of EMILLE have been successively refined. In the following subsections we describe the current 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 that faces any corpus builder is the identification of suitable sources of corpus data. Corpus design criteria for large corpora, while of use in guiding the search for corpus data, are of little use if no repositories of electronic text can be found with which to economically construct a corpus. This causes problems in Indic corpus building as the availability of electronic texts for Indic languages is limited. While this availability does vary by language, even at its best it cannot compare with the availability of electronic texts in English or other major European languages. As a result we were faced with the realization that much data that we would, in principle, like to include in our corpus existed in paper form only. On EMILLE, it would have been too expensive to pay typists to produce electronic versions of the sixty three million words of monolingual written corpus (MWC) data we wanted to collect. Even if the initial typing was affordable, checking the data for errors would have added a further cost. This cost would be increased further by the fact that tools which could have aid an error correction process, such as spell checkers, do not exist for many of the languages studied on EMILLE (McEnery & Ostler, 2000). Scanning in the text using an optical character recognition (OCR) program is a viable alternative to typing in printed text where languages are printed in the Roman alphabet. However, OCR programmes for Indic scripts are still in their infancy (for an example of some early work see Pal & Chaudhuri, 1995) and were not considered to be stable and robust enough for this project to use gainfully. We also wished to produce corpora in their native script and hence avoided romanized Indic texts altogether.

It is hardly surprising then that identifying sources of electronic data was at the top of our agenda from the very start of the project. As part of a pilot project to EMILLE[3], we ran a workshop that examined potential sources of such data for Indian languages. The workshop identified the Internet as being one of the most likely sources of such data. While we also considered publishers of Indic language books, religious texts, newspapers and magazines as a possible source of such data, the prevalence of old-fashioned hot-metal printing on the sub-continent made us realize early in the project that such

[1] Funded by the UK EPSRC, project references GR/K25267 and GR/M31699.
[2] Funded by the UK EPSRC, project reference GR/N19106
[3] This project, Minority Language Engineering, was funded by the UK EPSRC (Grant number GR/L96400).

sources were not likely providers of electronic data. Indeed a number of publishers expressed an interest in helping us, but none were able to provide electronic versions of texts that they had produced. In the light of the difficulty we experienced in gathering texts in their original scripts, we had to gather our corpus of MWC data from the web on the basis of four, largely pragmatic, criteria. The prime criterion for data collection in EMILLE was that for any text to be considered for inclusion in one of our monolingual written corpora it would have to be readily available as machine readable text in an Indic script. In real terms this meant that the material had to be gathered from the web. In gathering from the web, we elected not to use web-robots in order to gather the texts. The web texts we were looking at were rather complex. Adverts peppered the text. These adverts were often in a language other than that which we wished to gather (English adverts in Hindi texts, for example). Also, it was often the case that while individual stories on a newspaper page may change, the vast bulk of the page would remain the same. In electing to ignore the adverts and only select new material from each page we significantly complicated the retrieval task to the extent that we no longer found web robots useful for the task. We found it was faster for a human to visit the site, sort the text from the adverts, identify the useful material and save it. In doing so, as will be shown later, the human analyst was able to save the data with filenames that aided in the process of constructing the file header (see section 3.1).

The second criterion involved the format of the electronic text. 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. To date, we have yet to come across any Unicode data for Indic languages on the web. We found that creators of Indic documents on the internet typically rely on five methods for publishing texts online:

a) They use online images, usually in gif format. Such texts would need to be keyed in again, making the data of no more use to us than a paper version;
b) They publish the text as a pdf document. Again, this made it almost impossible to acquire the original text in electronic format. We were 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, formatting meant that words in texts would often appear in a jumbled order, especially when acquired from pdf documents that contained tables, graphics or two or more columns;
c) They use a specific piece of software in conjunction with a web browser. This was most common with Urdu texts, where a separate terminate-and-stay-resident program, such as Urdu 98, is often used to handle the display of right-to-left text;
d) 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;
e) 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 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 be virtually impossible.

The possible reasons for the bewildering variety of formats and fonts needed to view Indic language data on the web are many. However, the obvious explanation for the lack on 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 it was probable that those reading the text 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 or Courier, Indic fonts are not merely repositories of a particular style of character rendering. They represent a range of incompatible glyph encodings. To elucidate, in different English fonts, the decimal code 0042 is always used to represent the character “B”. However, in various fonts which allow one to write in Devanagari (Hindi) script, the hexadecimal code 0042 could represent a number of possible 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 & 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) in this font, two keystrokes would need to be entered. However, other fonts may use *He* a single keystroke to produce *He*.

We were also mindful that for every new 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 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 discussion of a new collaboration with a partner in India, the eventual corpus will now contain a wider range of genres (see below).

Our third criterion involved the amount of text we could collect from a single source. While we found numerous Indian language websites, not all of them were able to offer more than a few hundred words of data. The most useful sites were newspaper sites which provided daily updates, and usually contained archives that could be

exploited to gather yet more data. Therefore, we focussed on daily news websites for gathering the MWC data. In the absence of a wide range of data sources, which would produce many genres of texts in a corpus, newspaper data is useful as stories change from day to day, a number of writers contribute to the newspaper and within the newspaper a number of sub-genres such as news, politics, entertainment and sports can be identified.

Our final criterion was, in many ways, the most important. As the corpus will be publicly available, we had to obtain permission from the publishers of texts to use them. Fortunately, most of the online newspapers that we contacted were happy to let us include their texts in our corpus.

| Language | Millions of words |
|---|---|
| Assamese | 2.6 |
| Bangla | 5.4 |
| Gujarati | 7.8 |
| Hindi | 8.8 |
| Kannada | 2.2 |
| Kashmiri | 2.3 |
| Malayalam | 2.3 |
| Marathi | 2.2 |
| Oriya | 2.7 |
| Panjabi | 4 |
| Sinhalese | 4.9 |
| Tamil | 10.1 |
| Telegu | 4 |
| Urdu | 1.6 |
| Total | 60.9 |

Table 1. Word counts for each language in the EMILLE/CIIL Corpus as of April 2002

The four criteria in themselves would have allowed us to fulfil our original MWC project goals. However, over the past twelve months the MWC collection goals of the project have altered significantly. Thanks to a series of grants from the UK EPSRC[4] 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 with the Central Institute of Languages (CIIL) in Mysore, India with the goal of producing a wider range of monolingual written corpora than was originally envisaged on the EMILLE project. The effect of this change will mean that the uniform word counts of the monolingual written corpora will be lost. Each language will now be provided with varying amounts of data, though no language will be furnished with less than a million words. However, we will now be able to cover a much wider range of languages (14 rather than 7) and we will cover a wider range of genres. By a process of serendipity, the corpus data being provided by CIIL covers a wide range of genres other than newspaper material. The new EMILLE/CIIL corpus will, therefore, not only expand the range of languages of the final corpus, it will also extend the range of genres in that corpus[5]. Table one shows the state of the EMILLE/CIIL monolingual written corpora at present.

The collection phase for the EMILLE/CIIL MWC data is nearly finished, with only around 3 million words of data still to be collected. Consequently, the focus of the project is now falling increasingly on parallel and spoken data.

## 2.2. Parallel corpora

The problems we faced in collecting MWC data also faced us 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 the possibility of 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 translated from English originals: UK government health and advice leaflets. The leaflets we were able to gather were mostly in pdf format, though some also used a number of 8-bit encodings to represent Indic writing systems. 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 was economic to pay typists to produce Unicode versions of the texts using Global Writer, a Unicode word-processor which was able to handle the rendering of conjunctions, diacritics etc[6].

The research value of the British government data is very high in our view. The UK government is producing a large number of documents monthly in a wide range of languages. All of the texts are focused in areas which are term-rich, such as personal health, public health and social security. To build the parallel corpus we collected about seventy documents from the Departments of Health, Social Services, Education and Skills, and Transport, Local Government and the Regions. These documents have been translated from an English original into various languages. While we were only interested in Bangla, Gujarati, Hindi Panjabi and Urdu we found that many of these documents had also been translated into other languages including Arabic, Chinese, Polish, Somali and Vietnamese. Currently we are planning to expand the parallel data so that it covers these other languages. As the languages that are currently covered in the parallel data (Bangla, Gujarati, Hindi, Panjabi and Urdu) are all from the Indic branch of the Indo-European language family[7], the inclusion of additional languages could add other language families to the corpus. For example, Arabic and

[4] Grants GR/M70735, GR/N28542 and GR/R42429/01.

[5] The data provided by CIIL to the project covers a number of genres, including Ayurvedic medicine, novels and scientific writing.

[6] When the project began, Global Writer was one of the few word-processors which was able to handle the rendering of Indic languages in Unicode. Since then, Microsoft have made Word 2000 Unicode-compliant. However, unless running on a Windows 2000 machine the Unicode compliance of Word 2000 is not apparent.

[7] Although Urdu uses a radically different writing system to the others, as it is a modified form of Perso-Arabic rather than Sanskrit derived.

Somali are from different branches of the Afro-Asiatic family, Vietnamese is an independent language, Polish is from the Slavic branch of the Indo-European family and Chinese is from the Sino-Tibetan family. While a corpus of English texts translated into five Indic languages will undoubtedly be of use, increasing the typological diversity of the parallel corpus should enhance the worth of the corpus significantly.

Other than the need to type the data from paper copies, the parallel corpus also presents one other significant challenge: while most of the data we have access to 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 we are interested in. Our solution to this is to employ translators to produce versions of documents in the appropriate Indic language. While being 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 being non-official translations in their header.

### 2.3. Spoken corpora

For the collection of spoken data we have pursued two strategies. Firstly we explored the possibility of following the BNC 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 that were conducted in Gujarati for us, and a small number of the people who were involved in typing work on the project agreed to record their conversations with family and friends. The feedback that we received 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 Bangla 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 Channel as our sole source of spoken data. The BBC Asian Network readily agreed to allow us to record their programmes and use them in our corpus. The data source is excellent as it is broadcast on digital radio, hence ensuring high quality recordings. The five languages of the EMILLE spoken corpora are all covered by a phone-in programme. This programme is broadcast nightly for two hours, either in Bangla, Hindi, Gujarati, Panjabi or Urdu. The programme plays Indian music (which has 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 programme is apparent.

To date, we have banked sufficient data to construct our spoken corpora by sampling four weeks of radio programmes roughly once per quarter. We have now begun the process of transcribing the broadcasts and to date have transcribed 100,000 words of Urdu and 150,000 words of Bangla.

The orthographic transcription of the spoken data has thrown up two interesting issues, both, arguably, related to dialects. The first issue arose from the variety of Bangla spoken in the UK. Our main Bangla transcriber has lived in India for most of her life. She had no problems with transcribing conversations of other Bangla-speaking Indians, but when faced with tapes of the radio programme which featured Bangla speakers who lived in the UK, it became apparent that British-born Bangla speakers spoke a variety of Bangla rarely heard in India. UK Bangla speakers are overwhelmingly from the Sylhet region of Bangladesh and speak Sylheti, which one may either view as a separate language or a dialect of Bangla (Baker, Lie, McEnery & Sebba, 2000). As some of these words were unfamiliar to our non-Sylheti speaking transcribers, they were not transcribed. Instead the CES code <omit> has been used on such occasions e.g. <omit extent="1 syllable" cause="unclear dialect">. Our intention is that, at a later date, we will return to these points in the data with a Sylehti speaker and correct the transcription.

The second problem relates to prescriptive attitudes. As noted, the radio phone in data is of particular use as it means that a number of speakers are represented in the corpus, not all of whom are speakers of a nominal standard form of a language covered by EMILLE. This observation is not restricted to Bangla/Sylheti. It is apparent in all of the languages that we are gathering data for. This has caused some transcribers who have happily worked on typing parallel corpus data to refuse to work with the spoken material at all. They object to the representation of the Indic languages in the corpus. For example, one Hindi speaking transcriber from India refused to transcribe recordings of the BBC Asian Network Hindi radio programme, saying that linguists should only study 'classical Hindi texts and not the bastardised slang' that was used by South Asians living in the UK. Some of the differences that the transcribers have objected to relate to the code switching practices of the UK South Asian community. However, there are also objections to non-standard and non-prestige forms such as Sylheti being studied by linguists. While this is a manageable problem in the context of the EMILLE project, this experience served as a useful reminder that, while linguists may be happy studying all forms of language, for speakers of a specific language their willingness to help corpus builders may be influenced directly by their attitude to the forms of a language that a corpus linguist is seeking to represent and study.

## 3. CES encoding and conversion to Unicode

In this section we discuss aspects of text encoding and conversion which we are just beginning to work upon, having now collected a sizeable proportion of our corpus. In terms of corpus encoding, the texts are being marked up with header items and text elements viewed as essential in the Baker *et al* (1998) review of the corpus encoding needs of language engineers (e.g. elements to mark paragraphs, sentences, headings and foreign text). The corpus data is being annotated according to the Corpus

Encoding Standard recommendations[8], a set of minimal guidelines for the mark-up of corpora, compliant with the TEI. The CES is increasingly recognised as the standard for corpus building, with projects such as MULTEXT, PAROLE, BAF, TALANA and the American National Corpus project adhering to it.

### 3.1. The markup of MWC and parallel data

Our decision to collect material from the web was very useful as it furnished us with a fast track to CES compliance for MWC data. This data was collected initially in html format. This means that information placed in the document by the publisher and needed by CES, such as paragraphs, headings, line breaks and font face, size and colour was already encoded in each document. Also present was font information which is useful in determining sections of text that are encoded in different languages. For example, occasionally in the Indian language data, words appear which are in written in English. These words are encoded as <foreign lang="eng">. In short, the html code leaves us with only the <lang> and <s> elements to be included in each text in the MWC data.

As the MWC files are initially html files, the corpus texts already have short headers associated with them. While this header file is somewhat different to the CES header which we give each document, it does contain some of the same information, such as the time and date of the document created, the number of words in the document and the author. It is therefore relatively easy to automatically convert some of this existing information into a header that is compliant to CES. The additional information that we need for our header is initially encoded in the filename in which the data is initially stored. These file names then allow us to complete the header via an automated process. For example, take the file named tam-w-dinarkan-sports-07-01.00.htm. This filename gives us information about the language of the document, Tamil (tam), whether the text was originally spoken or written (w), the name of the online newspaper (Dinarkan), the genre (sports), and the date of publication (January 7th 2000). These fields are easily inserted into the header, completing the header for each MWC file.

As the parallel data was being typed in the appropriate CES markup was introduced to the text. While this did not cause a significant increase in work for the typists, it did create an additional overhead in checking the data to ensure that the markup guidelines had been applied consistently by the typists. Sometimes differences were found, but often these had been caused by differences in the translated forms of the documents, e.g. a bulleted list in English being represented as a paragraph in a Hindi translation. Where differences in markup across the parallel files accurately reflect differences in original documents, the inconsistencies have been left in the corpus. Where these inconsistencies have been caused by inconsistent/inaccurate application of the CES guidelines, we have sought to correct the markup.

### 3.2. The markup of spoken corpus data

Unfortunately, the CES had not published guidelines for the annotation of spoken texts when we began the project, so we have complied our corpora using the TEI guidelines for spoken annotation instead. Utterances are encoded using the format <u id="x" who="xxxxx"> and each speaker is referred to with a unique five digit code. For example, BM300 stands for Bangla Male, number 300.

### 3.3. Font harmonization and GATE

With regard to converting the numerous 8-bit fonts to Unicode, the main issues we have encountered so far have centred around the rendering of conjunct and half-form characters and diacritic vowels. In spite of limiting the range of data sources we have tapped in order to limit the font conversion task, to date we have come across 23 separate encodings of the 7 writing systems we are concerned with in this project. This is further complicated by the fact that each file contains at least two writing systems; the script of Indic text and the roman alaphabet which is used in the CES encoding tags and can appear in short passages of English embedded in the text. It may also be the case, for example, that a quote in Hindi written in Devanagari may appear in the middle of a Panjabi text written in Gurumukhi. Consequently any conversion program needs to be markup aware, and needs to be able to interpret <lang> elements in the text in order to work appropriately. Such a conversion programme is currently under development on the EMILLE project. The program in itself may appear trivial, but it hides the Sisyphean task of actually gathering together the various 8 bit encodings of these language in order to construct a robust mapping programme that makes the task rather open ended and hence difficult.

A difficulty we had to address before even starting the font conversion process was the development of an LE environment in which to carry out the work. While the decision not to produce the corpus in a legacy 8 bit encoding was forward looking, it led to a significant problem within the project and presents a challenge to future users of the corpus: how can one work with Unicode corpora? The response of the EMILLE project to this problem has been to work on the development of a Unicode compliant version of the General Architecture for Text Engineering (GATE, Cunningham et al, 2000). In part, this work was assisted by porting the latest version of GATE to Java, which provides some facilities for working with Unicode. However, GATE's capacity to allow users to work with Unicode was extended beyond that provided by Java in three ways[9].

Firstly, GATE now has a Unicode compliant editor with input methods for many languages. This editor uses a virtual keyboard window with the characters of the language assigned to the keys on a standard keyboard. Data can then be input either by typing as normal, or with mouse clicks on the virtual keyboard.

Secondly, in order for the editor and other programs requiring input to work appropriately, GATE now allows the user to select an input language. If not chosen, by default GATE will choose a Unicode font if it can find one on the platform a user is employing. If this is not appropriate the capacity exists for a user to select another input method, for example an 8-bit font, where needs be.

[8] See http://www.cs.vassar.edu/CES/

[9] See for full details of the newest version of GATE see http://gate.ac.uk/sale/tao/index.html#x1-550002.26.

Finally, it would be rather clumsy if a user had to select a language and font every time they wished to simply view a file. Consequently, on looking at a text GATE will initially default to the default encoding on the users machine, on the assumption that this will often times be the right encoding for any given file a user may wish to look at. Then users only need to reset this parameter as and when needed.

With these three important developments in GATE we had a platform on which our corpus data could be viewed in Unicode or any of its 8 bit font encodings.

The provision of an environment in which the data can be explored and manipulated has not merely paved the way for work on font conversion, it has also acted as a spur to our work on language engineering tools for Indic languages.

## 4. Development of LE tools

To date most of our effort in the area of LE tools on EMILLE has fallen into two areas – the development of a Unicode compliant sentence aligner and preliminary work on the development of a part-of-speech (POS) tagger for an Indic language. In this section we will focus exclusively on the development of the POS tagger, as the work on the aligner is focused on developing input methods rather than novel research as such.

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 the POS tagging work 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 Urdu 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 in order to 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 the 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[10] 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,[11] 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;[12] the distinction between 'marked' and 'unmarked' nouns; the Urdu honorific pronoun āp, which does not fit easily into any of the EAGLES categories for pronouns; and the borrowed Persian enclitic called izāfat. However, the idiosyncrasy of Urdu which is most illustrative of this issue is the "zimmah dār problem".[13] None of these problems were insurmountable. EAGLES has proved a robust and useful framework within which to approach Urdu POS tagging.

## 5. 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, some changes have occurred on the project which have been responses to unexpected factors,

[10] 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.

[11] 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 irregular auxiliary elements.

[12] The participles and the infinitive can all display case.

[13] The *zimmah dār* problem is so called because it was first encountered during an attempt to manually tag some sample sentences using an early version of the tagset. The word *zimmah dār*, "responsible", was immediately obvious as problematic. Other examples include *Tēlī fōn*, "telephone"; *xūb tar*, "better". The problem is common in borrowed vocabulary (in these cases, from English and Persian).

such as the reluctance of members of the minority communities to engage in the recording of everyday spontaneous speech. With regard to the LE tools produced by EMILLE the greatest contribution of the project to date has been to the on-going development of GATE, specifically in the area of Unicode compliance. However, in the near future further resources such as a part-of-speech tagger for Urdu, font conversion software and Unicode sentence aligners will become available.

## 6. References


Baker, J. P., Burnard, L., McEnery, A. M. & Wilson, A. (1998), Techniques for the Evaluation of Language Corpora: a report from the front. *Proceedings of the First International Conference on Language Resources and Evaluation*: Granada.

Baker, J.P., Lie. M., McEnery, A.M. and Sebba, M. 2000. Building a Corpus of Spoken Sylheti. Literary and Linguistic Computing, 15:419-431.

Bureau of Indian Standards, Indian Standard Code for Information Interchange, IS13194, 1991.

Crowdy, S. 1995. The BNC spoken corpus. In G. Leech, G. Myers and J. Thomas (eds.), Spoken English on computer: transcription, mark-up and application. Longman:London.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. and Wilks, Y. 2000. Experience of using GATE for NLP R&D. In *Proceedings of the Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000*:Luxembourg.

Leech, G. and Wilson, A. (1999) Standards for tagsets. In van Halteren, H. (ed.), *Syntactic Wordclass Tagging*. Kluwer:Dordrecht.

McEnery, A., Baker, J.P. and Burnard, L. 2000. 'Corpus Resources and Minority Language Engineering', in M. Gavrilidou, G. Carayannis, S. Markantontou, S. Piperidis and G. Stainhauoer (eds) *Proceedings of the Second International Conference on Language Resources and Evaluation*:Athens.

McEnery, A.M. & Ostler, N. 2000. 'A New Agenda for Corpus Linguistics – Working With All of the World's Languages', Literary and Linguistic Computing, 15:401-418.

Oksefjell, S. 1999. A Description of the English-Norwegian Parallel Corpus: Compilation and Further Developments. International Journal of Corpus Linguistics, 4:197-219.

Pal, U. and Chaudhuri, B.B. 1995. Computer recognition of printed Bangla script. International Journal of System Science, 26:2107-2123.

Schmidt, R. L. 1999. *Urdu: an essential grammar*. Routledge:London.

Singh, S., McEnery, A. and Baker, J.P. 2000. Building a Parallel Corpus of English/Panjabi. In J. Veronis (ed.), Parallel Text Processing, Kluwer:Dordrecht.