EMILLE: Building a corpus of South Asian languages.

Anthony McEnery, Paul Baker, Rob Gaizauskas*, Hamish Cunningham*
Dept. Linguistics, Lancaster University, Bailrigg, Lancaster, LA1 4YT
*Dept. Computer Science, Sheffield University, Sheffield, S1 4PD

Abstract

The paper describes the goals of the EMILLE[1] (Enabling Minority Language Engineering) Project at the Universities of Lancaster and Sheffield. Building on the findings of MILLE (the Minority Languages Engineering project[2]), EMILLE is focusing upon problems of translating 8-bit language data into Unicode, and is working towards a solution based around the LE (language engineering) architecture GATE[3]. A description of ongoing work on constructing the 63 million word EMILLE corpus of spoken and written data is also given. Our goal is to provide the basic architecture and data required to encourage research into South Asian language engineering. In particular, this will support the development of translation systems and translation tools which will be of direct use to translators dealing with languages such as Bengali, Hindi and Panjabi both in the UK and internationally.

Introduction

Corpus building in Europe has traditionally focussed on languages that are indigenous to European countries: English, French, Spanish, German, Italian[4] etc. Users of such languages benefit from an extensive range of computational resources: fonts, word-processors, spell-checkers, online dictionaries, thesauri and automatic translation utilities. However, in the UK there are sizeable communities of speakers of non-indigenous minority languages (NIMLs): e.g. Bengali, Cantonese, Gujarati, Panjabi, Urdu, etc. Estimating the number of speakers of these languages in the UK is difficult; for example there has never been a "language" question in the British Census. Yet the communities are undoubtedly large. The 1991 census included a question on "ethnic origin" for the first time. This showed that there are 840,255 Indian residents, 476, 555 Pakistani residents and 162, 835 Bangladeshi residents in the UK (Peach 1996: 11). While, ethnic origin should not be viewed as being directly translatable to the number of speakers of a language, there is evidence to show that within the communities of South Asian immigrants to the UK, multilingualism is widespread. Crucially, the use of languages from South Asia is continuing in these communities, even amongst members of the community born and raised in the UK. Consequently, it is fair to estimate that the speakers of languages from South Asia in the UK numbers in the hundreds of thousands. This claim is entirely in line with earlier more focussed on language practices in specific areas of the UK such as the Linguistic Minorities Project (Morawska and Smith, 1984) and the Inner London Education Authority Language Census (Alladina and Edwards, 1991) which found considerable evidence to support the assertion that there exist large communities of NIML speakers in the UK. The existence of these communities means that the domestic translation market in the UK is currently – and is likely to be for the foreseeable future – focused around South Asian languages. However, the computational resources for these languages are scant (Somers 1997: 3-7). Our work is oriented towards providing the means of redressing this imbalance, and enabling

language engineering on UK NIMLs, with domestic translation in the UK being seen as an important long term beneficiary of our work. Our focus on UK NIMLs has been informed and refined by undertaking a review of NIML language engineering in the UK as part of the MILLE project.

The MILLE pilot project

The MILLE (Minority Language Engineering) project was an EPSRC-funded 18-month research project at Lancaster University, designed to investigate the development of corpus resources for UK NIMLs. Its main goals were to determine:

1. the feasibility of constructing NIML corpora
2. the extent and availability of existing NIML data
3. the needs of NIML-speaking communities
4. the requirements of language engineers who will need tools in order to exploit corpora
5. methods of putting the data into machine readable, accessible format

As part of this study a major review of the needs of the language engineering (LE) community was carried out, with over 65 research centres worldwide responding. The results of the review revealed that many researchers wanted to work with Indic[5] languages (Baker & McEnery, 1999). However, a lack of corpus resources, problems involving data interchange and the need for a LE architecture capable of supporting work in such writing systems means that this type of work is unlikely to be carried out at present.

As one of the main goals of the MILLE project was to investigate the *feasibility* of building NIML corpora (goal 1), a good part of the project concentrated on issues surrounding the location of language data sources (goal 2), conversion of corpus data to a standardised electronic form, the application of mark-up schemes to Indic language data, and parallel text alignment (goal 5). In this paper we will focus on goal 5 to illustrate the kinds of issues we uncovered with relation to NIML corpus building[6]. Our investigation of this goal was informed by the construction of a number of small sample corpora in Chinese, Panjabi and Sylheti[7]. These corpora were successfully annotated using the Text Encoding Initiative (TEI, Sperberg-McQueen & Burnard, 1994) guidelines for electronic encoding and interchange (Singh McEnery & Baker, 2000), convincing us that corpus encoding as such is not necessarily a major issue for these languages. The major technical issue we encountered related to the multiple encoding formats used to represent text in Indic languages especially.

Encoding Problems

Several problems relating to character encoding were uncovered as a result of building our sample corpora. While existing electronic data was difficult to obtain, it was not impossible to gather, even though some compromise had to be made in collecting data that was representative of all genres. The main difficulty in building Indic NIML corpora is the need to standardise the language data into a single character set. Most Indic languages are represented electronically with 8-bit fonts. However, while ISCII (Indian Standard Code for Information Interchange) provides a standard 8-bit code table and keyboard layout for Devanagari (the writing system used to represent Hindi), many font creators use different keyboard layouts and non-standardised character sets. When collecting data from multiple sources,

standardisation becomes a problem, as unlike in English where "a" (or hexadecimal character 0061) will always be "a", whether it is rendered in Times New Roman, Arial or Courier fonts, the code 0061 could be used in different Devanagari fonts to represent a whole range of characters.

A solution has been proposed involving a 16-bit "universal" character set[8]: the Unicode Standard, which ultimately aims to use a single character set to represent all languages. Within Unicode characters are encoded via script rather than language and the Devanagari Unicode table is based on a template from the 1988 version of the ISCII standard. In essence each Unicode character receives a 4 digit hexadecimal number, which is standardised - so a number which is assigned to a character will only ever signify that character. Writing or viewing Unicode, above the first 256 characters (which are reserved for English and punctuation), however, is problematic. Without an ability to render a large range of writing systems, some software systems which claim to be Unicode compliant are unable to display some writing systems: in reality they are compliant to UTF-7 or UTF-8[9]. Windows NT, for example claims Unicode-compliance but doesn't have built-in support for the whole multilingual character set. Similarly, the Java Developer Kit has the potential, but not the available font rendering software, to handle Unicode data. Finally, Netscape Communicator 4.0 allows Unicode documents to be browsed in UTF-7 and UTF-8 but a multilanguage support plug-in must be installed to go any further than this (and this plug-in only handles a subset of Unicode scripts).

In the MILLE project, we collected data that had been created using 8-bit fonts, and used UniEdit, an editor from Duke University to convert the text to Unicode. For each font we encountered we had to create a different interchange table that was implemented by a conversion tool specific to UniEdit. In this way we were able to encode most of our data in Unicode. However, there were several problems inherent with this method. First, UniEdit lacked some of the Unicode characters needed for encoding Indian languages. Secondly, and most importantly, problems arose because the Unicode Standard contains a number of rendering rules for Indian languages which must be applied by a font rendering engine for the resultant text to be displayed appropriately. In Devanagari, for example, characters can combine in ways which create other characters, that are not represented in the Unicode Standard. The Standard notes "*in a font that is capable of rendering Devanagari, the set of glyphs[10] is greater than the number of Devanagari Unicode characters*" (Unicode Consortium 6-38).

For example, according to Unicode's rendering rules, typing in the sequence: ह् र should result in ह्र being displayed upon the screen. Most word processors that use 8-bit fonts do not use such rendering rules. Instead they simply list all (or many) of the possible characters, including a large range of diacritic symbols. So when using an 8-bit font with a standard word processor, to achieve the sequence ह्र on the screen, the character for ह followed by ◌ं will be typed. However, because the diacritic ◌ं only occurs in the Unicode Standard when other characters are combined together, it does not exist as a "real" character entity. Instead the Unicode Standard provides the rules and then leaves it up to software designers to implement them correctly. We encountered a major problem in our use of UniEdit as it does not implement these rules.

Another Unicode editor, Global Writer *does* employ these rules, but a further problem was encountered when we tried to transform 8-bit files for use in Global Writer. As Global Writer interprets all of Unicode's rendering rules it requires characters to be entered in a particular sequence. A diacritic vowel, e.g. ਿ, which is attached to a consonant, e.g. ਠ , *must* always be entered after the consonant character. So to create the sequence ਠਿ on-screen, the characters must be entered as ਠ followed by ਿ. But with an 8 bit font, this sequence may be entered as ਿ followed by ਠ . So resequencing of the 8 bit data stream may be necessary if it is to be made compliant with the rendering rules of a Unicode compliant editor. A solution is required to the problem of interpreting conflicting 8-bit font representations of South Asian writing systems, converting them all into standardised Unicode characters, implementing the necessary rendering rules and resequencing the order of the characters so that the text still makes sense in Unicode. At present no Unicode editor is capable of this[11]. As such, this represents a major impediment to the construction of corpora of South Asian languages. Issues such as this together with the general need for South Asian corpus data in language engineering were the prime motivations for the EMILLE project.

EMILLE: aims

In order both to produce a framework for language engineering for South Asian languages and to generate corpus data to enable such work, the EMILLE (Enabling Minority Language Engineering) was funded by the ESPRC. EMILLE is a joint project between the Universities of Lancaster and Sheffield. The project has three main goals: to extend an LE architecture, to build corpora of South Asian languages and to develop basic LE tools.

Goal 1 - extend an LE architecture

The project is establishing an LE architecture within which minority LE may take place. To be truly generic platforms, LE architectures cannot be limited to specific languages/writing systems. A recent EPSRC workshop on LE architectures[12] led to a conclusion that LE architectures need to expand beyond their current focus on European languages. To this end, EMILLE is implementing a UNICODE compliant version of GATE, the General Architecture for Text Engineering (Cunningham *et al*, 1997; Gaizauskas *et al*, 1996), a widely used platform for the development and reuse of LE components. The system is a software architecture/development environment that supports researchers in natural language processing and computational linguistics, as well as developers who are producing and delivering LE systems. It has been used for a wide variety of applications including information extraction (Gaizauskas and Wilks, 1998) and sense tagging (Cunningham, Stevenson, and Wilks, 1998).

In EMILLE we are implementing tools within GATE to cope with font mapping to allow our corpora to be standardised around UNICODE. GATE is also being adapted to allow it to implement the UNICODE standards for conjunct formation. Existing alignment software is being embedded and evaluated on Indic languages in GATE. The corpus validation tools recommended by Baker *et al* (1998b) are being incorporated within GATE, and basic tools developed to allow for the rapid development of corpus headers and mark-up. With corpus building and validation

tools in place, GATE will be an architecture within which TEI conformant corpus texts can be developed and validated.

Goal 2 - develop corpora

EMILLE is generating written language corpora of at least 9,000,000 words for Bengali, Gujarati, Hindi, Panjabi, Singhalese, Tamil and Urdu. These are the Indic languages indicated as being those most wanted by the LE community in the Baker & McEnery (1999) survey. For those languages with a UK community large enough to sustain spoken corpus collection (Bengali, Gujarati, Hindi, Panjabi and Urdu) EMILLE is also producing spoken corpora of at least 500,000 words per language.

Written Data

We are in the process of collecting 200,000 words of parallel text. The remainder of the text collected is monolingual corpus data. We have chosen a figure of 200,000 words, as a corpus of this size produced by the MULTEXT project proved an adequate basis for the largest comparative evaluation exercise for alignment tools yet undertaken (Langlais *et al* 1998). Data donors in the UK and South Asia are providing the parallel corpus data.

The monolingual texts are being gathered both in the UK and from the Indian subcontinent. The corpora will contain a minimum of 20% of texts gathered from UK sources. Contacts established on the MILLE project, such as Lake House printers in Sri Lanka[13], The Dept. of Health, the Sikh Parliament in Birmingham and community newspapers in the UK are being used to gather the data. Several online Indian newspapers have also given permission to include text from their websites in our corpus. We also have permission to use health leaflets, religious texts and novels. It is our intention to make the corpus as representative of as many domains of writing as possible.

There is also a small project, designed to support EMILLE, funded by the British Council, which is building a network of academics interested in corpus building in South Asia. This network is also contributing to the data collection effort, with further contacts being made as the project proceeds.

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* (1998b) review of the corpus encoding needs of language engineers, with the addition that country of origin for each text in the corpus is being encoded. The corpus data is being annotated according to the Corpus Encoding Standard recommendations[14], 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 implementing CES guidelines.

Spoken Data

The spoken corpus data is being gathered from communities across the UK. A collection model established in Lie *et al* (1999) is being followed in order to gather data in ways that vary the amount of code switching to be expected, a critical variable in this sort of data collection. We are using community links established on the MILLE project to gather the data. Again, where necessary, wider community contacts are being used to recruit more informants. The corpus data is being gathered on mini-disks. The digitised sound wave of the minidisks will be stored and released as part of

the final project deliverables. This use of digital media to collect the data will ease the transfer of the data to computer. The data is also being transcribed[15]. Estimates for the production of the transcriptions are based upon previous experience with MILLE. The estimate is prudent, and includes time for training transcribers. Transcription will occur in the native script of the speakers where possible. Throughout, regular checking of transcriptions produced is being undertaken by analysts specifically employed to carry out random quality checks.

The metadata gathered to accompany each transcription are limited to age, gender and occupation. These are objectively verifiable categories. Categories such as social class, which may appear attractive, are subjective and unreliable.

We have also obtained permission to transcribe spoken data from a number of radio and television stations which broadcast programmes in Indic languages. Therefore the spoken section of our corpus will contain examples of spontaneous and scripted speech.

Goal 3 - develop basic LE tools

As noted above we are developing tools within the GATE framework to allow for mapping a diverse range of font-based representations of Indic writing systems into UNICODE. However, the project is also undertaking the part-of-speech tagging of Urdu in both spoken and written form and is adapting existing alignment software to sentence align the parallel corpora within EMILLE. These tools will be embedded within the GATE architecture. For both tagging and alignment, EMILLE is drawing upon existing techniques in non-Indic language engineering and it is one of the research objectives of the project to investigate the extent to which these techniques can be adapted to meet the needs of Indic LE.

The Tagger

Part-of-speech tagging is one of the most common and useful forms of corpus annotation used by language engineers (Leech, 1997). In order to enhance the usefulness of the corpus resources generated by EMILLE, it was decided to develop a part-of-speech tagger for Urdu, with the aim of part-of-speech tagging the spoken and written Urdu corpus data. The work on this tagger is proceeding in 6 stages:

1. Develop a part-of-speech tagset. Throughout this process, the EAGLES[16] part-of-speech tagset standards are proving to be a useful guide, though the strong possibility exists that we will have to refine those standards to apply to non-European languages.
2. Develop a tokenizer and embed it in GATE.
3. Select a tagger for training. A range of tagging technologies available within GATE are being reviewed, in order to determine which is likely to be the most successful for the task in hand. Several taggers are available within GATE, e.g. Brill (1992), CLAWS (Garside, Leech & Sampson, 1986) and ROBOTAG[17].
4. Manually tag a suitable sample of text for training the tagger.
5. Begin tagger training. The tagger chosen in phase two will be trained to work on both written and spoken language files.
6. Tag the written and spoken language corpora.

While at an early stage, we aim to complete all six stages of this tagger development in time to release the tagger into the public domain by the end of the project.

Alignment

The alignment undertaken on the project will be carried out using the McEnery & Oakes (1996) version of the Gale & Church (1993) sentence alignment algorithm, which has already been used successfully on Panjabi (Singh *et al*, 2000). Evaluation of the success of the alignment will be undertaken by native speakers. Existing sentence alignment software will be used to conduct an alignment competition similar to that of Langlais *et al* (1998a,b). However, as we are over two years away from the possibility of such a competition on Indic languages, we can only guarantee at this stage that the parallel corpus will be sentence aligned using the McEnery & Oakes aligner and that an evaluation of the alignment will be undertaken. The alignment will follow the evaluation procedure set out in Langlais *et al* (1998a,b) and take into account work such as Somers (1998).

Conclusion

EMILLE is being undertaken because the creation of a language engineering architecture for Indic languages, populated with basic LE resources and tools, will provide the basis for further research activity related to these languages. Without the architecture, constructing the corpora and re-using existing tools would be difficult. Once the architecture has allowed basic resource and tool construction, the architecture, incorporating the corpora and tools, will form a sound basis for further language engineering and resource development. EMILLE is establishing the foundations of Indic language engineering and should be of major significance for the development of translation systems and aids designed to support translation for the UK domestic market.

Notes

[1] Funded by the UK EPSRC, project reference GR/N19106.

[2] Funded by the UK EPSRC, project reference GR/L96400.

[3] Funded by the UK EPSRC, project references GR/K25267 and GR/M31699.

[4] E.g. The British National Corpus, The Crater Corpus of Spanish, French and English, the ET10-63 Corpus of English and French, the MULTEXT Corpora of English, German, Italian, Spanish and French.

[5] A term we will be using in this paper to refer to the languages of South Asia. As such it is an umbrella term, covering a range of Dravidian, Indo-Aryan and Tibeto-Burmese languages. EMILLE, however, is concerned with a sub-set of Dravidian and Indo-Aryan languages only.

[6] Readers interested in our exploration of the other issues raised here, and particularly our rationale for focusing on Indic languages in the EMILLE project, should see McEnery, Baker & Burnard (2000).

[7] A dialect of Bengali, spoken by about 95% of Bangladeshis living in the UK (Lie et al 1999).

[8] With the 32 bit version ISO-10646.

[9] Unicode Transformation Format.

[10] In Unicode glyphs are representations of individual characters. So a character may be "a" but it may have numerous glyphs e.g. "a", "a" or "a".

[11] For a fuller description of the encoding problems associated with building South Asian language corpora see Baker et al (1998a).
[12] "A Workshop on Language Processing Architectures and the Use and Distribution of Language Resources", EPSRC ref. GR/M44545.
[13] For example, Lancaster established an agreement with Lake House to provide data to EMILLE on the MILLEFT project.
[14] See http://www.cs.vassar.edu/CES/
[15] By transcription here we mean broad orthographic transcription.
[16]The Expert Advisory Group on Language Engineering Standards. See: http://www.ilc.pi.cnr.it/EAGLES96/home.html
[17] A rule driven tagger under development at Sheffield.

References

Alladina, Safder & Edwards, Viv. (1991), *Multinglingualism in the British Isles*. London: Longman.

Baker, P. & McEnery, A. (1998), *Needs of language-engineering communities; corpus building and translation resources*. MILLE working paper 7, Lancaster University.

Baker, P., Burnard, L., McEnery, A. & Sebba, M. (1998a), *Beyond the 8 bit character set: the representation and exchange of Indian and Chinese corpus data*. MILLE working paper 2, Lancaster University.

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

Brill, E. (1992), "A simple rule-based part-of-speech tagger", *In Proceedings of the Third Conference on Applied Natural Language Processing*.

Cunningham, H., Humphreys, K., Gaizauskas, R., and Wilks, Y. (1997), Software Infrastructure for Natural Language Processing. *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington DC.

Cunningham, H., Stevenson, M. and Wilks, Y. (1998), "Implementing a Sense Tagger within a General Architecture for Language Engineering", *Proceedings of the Third Conference on New Methods in Language Engineering (NeMLaP-3)*, Sydney, Australia.

Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P. and Humphreys, K. (1996), "GATE -- an Environment to Support Research and Development in Natural Language Engineering", in *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-96)*.

Gaizauskas, R. and Wilks, Y. (1998), "Information Extraction: Beyond Document Retrieval", *Journal of Documentation, 1*.

Gale, W.A., & Church, K.W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", *Computational Linguistics* 19:1.

Garside, R., Leech, G. And Sampson, G. (1986), *The Computational Analysis of English*, Longman: London.

Langlais, P., Simard, M., & Véronis, J. (1998), Methods and practical issues in evaluating alignment techniques. *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistic* , Montréal, Canada, 1998.

Langlais, P., Simard, M., Véronis, J., Armstrong, S., Bonhomme, P., Débili, F., Isabelle, P., Souissi, E., & Théron, P. (1998), "ARCADE: A co-operative research project on bilingual text alignment" in the *Proceedings of the First International Conference on Language Resources and Evaluation (LREC),* Granada, Spain, 1998.

Leech, G.N. (1997), "Grammatical Tagging", in R. Garside, G. Leech & A. McEnery (eds) *Corpus Annotation*, Longman: London.

Lie, M., Baker, P., McEnery, A. & Sebba, M. (1999), "Building a Corpus of Spoken Sylheti", in N. Ostler (ed) *The Proceedings of the 3rd Conference of the Foundation for Endangered Languages*. Foundation for Endangered Languages, Bath.

Morawska. A, and Smith, G. (1984), *The Adult Language Use Survey of the Linguistic Minorities Project. The Data in Context*. LMP/CLE Working Paper No. 9. Institute of Education: London.

McEnery, A., Baker, J. & Burnard, L. (2000), "Corpus resources and minority language engineering". In *Proceedings of LREC 2000*, ELRA, Paris.

McEnery, A. & Oakes, M. (1996), "Sentence and Word Alignment in the CRATER Project". In *Using corpora for language research : studies in honour of Geoffrey Leech*. Thomas, J. & Short, M. (eds). Longman: London. ed. by Jenny Thomas and Mick Short.

Ostler, N. (1999), "Language technology and the Smaller Language", *ELRA Newsletter,* 4 (2).

Peach, C. (1996), *Ethnicity in the 1991 Census. Volume Two. The ethnic minority populations of Great Britain*. HMSO: London

Reynolds, M. (1996), "Punjabi/Urdu in Sheffield: are the languages being kept or lost?", *Language Issues* 8 (1).

Singh, S., McEnery, A. & Baker, P. (2000), "Building and Aligning a corpus of Panjabi-English", in J. Veronis (ed*) Parallel Text Processing*, Kluwer Academic Publishers: Dordrecht.

Somers, H. (1997), "Machine Translation and Minority Languages", *Translating and the Computer 19: Papers from the Aslib conference,* London.

Somers, H. (1998), "Further Experiments in Bilingual Text Alignment", *International Journal of Corpus Linguistics*, 3 (1).

Sperberg-McQueen, C.M. & Burnard, L. (1994), Guidelines for electronic encoding and interchange (TEI P3). Chicago and Oxford: Text Encoding Initiative. cf. http://www.uic.edu/orgs/tei/p3/doc/p3.html

Unicode Consortium. (1997), *The Unicode Standard, Version 2.0*. Addison-Wesley: New York.