

Corpus Resources and Minority Language Engineering

Tony McEnery, Paul Baker and Lou Burnard*

Department of Linguistics, Lancaster University
Bailrigg, Lancaster, LA1 4YT, UK
mcenery@comp.lancs.ac.uk

*Oxford University Computing Services,
13 Banbury Road, Oxford,
OX2 6NN, UK

Abstract

Low density languages are typically viewed as those for which few language resources are available. Work relating to low density languages is becoming a focus of increasing attention within language engineering (e.g. Charoenporn, 1997, Hall and Hudson, 1997, Somers, 1997, Nirenberg and Raskin, 1998, Somers, 1998). However, much work related to low density languages is still in its infancy, or worse, work is blocked because the resources needed by language engineers are not available. In response to this situation, the MILLE (Minority Language Engineering) project was established by the Engineering and Physical Sciences Research Council (EPSRC) in the UK to discover what language corpora should be built to enable language engineering work on non-indigenous minority languages in the UK, most of which are typically low-density languages. This paper summarises some of the major findings of the MILLE project.

1. Introduction

Corpus data is the *sine qua non* of many modern language engineering applications. It follows, therefore, that where corpus data for a language is lacking, the ability of language engineers to generate tools/systems for use with that language is seriously reduced. A lack of corpus data for a language may have severe consequences for the future of that language, for as Ostler (1999:3) says “languages which do not take a full part in the electronic media are doomed to stagnate, if not atrophy”. This paper outlines how we have developed our strategy for working on South Asian languages. We argue, based upon our work to date, that a real need exists in the language engineering community for corpus data in these languages. This paper is about why such corpora should be built, and how we intend to build them. As such the paper has three main goals. Firstly, we will review the state of language processing technology for low density¹ languages, building upon the work of Somers (1997). Secondly, we want to present the findings of a major review (with over 60 research centres world-wide participating) of the needs of language engineers and translators in relation to low density languages. Finally we will present our response to the problems we outline and the needs we uncovered. In doing so, we will present research just beginning at Lancaster and Sheffield Universities to extend a current language engineering architecture, GATE, to act as an architecture for low-density language engineering.

¹ Note we will use the terms low density language engineering and minority language engineering interchangeably in this paper, as both terms have currency in the corpus linguistics and language engineering communities.

2. The State of Minority Language Engineering

Before we present the survey of needs that we conducted, we would like to review, briefly, the current state of play in minority language engineering. Such a review clearly underlines why researchers are interested in gathering corpus data and building corpora for low density languages - the state of the art is shockingly poor. Somers (1997:6) gives a table which lists the availability of various resources for different “exotic” languages. Somers’ table shows a disappointing lack of resources except in the case of word processors and fonts, with Chinese, Greek, Polish and Arabic being the best provided for.

The computational landscape changes rapidly, however, and an updated version of the review, carried out in 1999 shows a less gloomy state of affairs (Baker and McEnery, 1999). Yet the situation is still relatively bleak for Indic² languages, especially Sylheti, which in its written form can reasonably be viewed as an endangered language (Lie et al, 1999). To highlight some of the problems faced by minority language engineering, we will consider four areas in slightly more depth here: word processing, UNICODE, dictionaries/term banks and OCR software.

2.1. Word processing

It is possible to find font-based representations of a wide range of writing systems now. A good font-based solution, used with a word processor such as Microsoft Word, is the most common solution employed to allow users such as translators to word process, for example, Indic languages. However, working with font-based solutions can be problematic. Initially there is a learning

² A term we will be using in this document 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.

overhead involved in using them – one needs to remember how a Roman keyboard maps to another writing system³. Another problem is common to all font-based solutions - different font-based representations of the same writing system use different code tables/keyboard mappings. For example, pressing a letter ‘a’ on a Roman keyboard mapped to the Gurumukhi writing system using fonts available from *Punjabi Resources on the Web*⁴ will produce the character ‘a’, while the same key press using the Gurbani Lipi font for Gurumukhi will produce ‘a’. Font-based solutions of this sort are just not the way for language engineering to go, as even if one were to try to harmonise language engineering around one set of font-based solutions, the existing font-based solutions all seem to have rather idiosyncratic shortcomings. For example, the fonts for Indic languages that we have reviewed provide a patchy solution for diacritics and conjunct characters (Singh, McEnery and Baker, 2000). In short, the existing solution for word-processing in low density languages using a font-based solution leads to a chaotic and often unsatisfactory range of solutions to the basic problem of being able to wordprocess in a most rudimentary fashion.

Moving beyond such simple word processing reveals further problems. For example, few language-specific or multilingual word-processors are available which include menus and help in multiple languages. Basic functions of wordprocessors, long accepted as standard for wordprocessors of European languages, such as spell checkers, are rarely truly available for Indic and other low density languages. With such poor provision of basic language engineering products, it is hardly surprising that such specialist resources as electronic dictionaries and termbanks are rare. However, before discussing the provision of basic electronic resources, it is worth considering the potential role of Unicode in overcoming some of the problems outlined in this section.

2.2. Unicode

Unicode is perceived by many language engineers as the future for multilingual text encoding (Baker, Burnard, McEnery and Wilson 1998). It is envisaged that before very long all electronic text will be formatted in Unicode, alleviating the need for the font-based solutions currently used for many writing systems, and hence removing many of the problems outlined in section 2.1. It is, therefore, important that the writing systems of low density languages are fully represented in Unicode. Those writing systems which are not included may find themselves placed at a permanent encoding disadvantage in the future.

At the time of writing, Unicode 3.0 has just been released and a number of significant additions to the writing systems covered by Unicode have been included (for example, Cherokee, Burmese, Ethiopic, Maldivian, Singhala, Khmer and Yi). Even so, there are writing systems which Unicode has not yet addressed, such as Nagri, the writing system of Sylheti. Hence Unicode, while a welcome initiative, still has to expand to include the fullest range of low density languages possible.

³ Producing a keyboard overlay is another solution.

⁴ <http://theory.tifr.res.in/bombay/history/people/language/punjabi.html>

2.3. Dictionaries and term banks

Electronic dictionaries can be useful for language engineering – electronic dictionaries containing, for example, semantic field, part-of-speech and pronunciation information have been available for languages such as English for over a decade now, and have been used in a range of language engineering applications. However, such resources are thin on the ground for a wide range of low density languages, and on closer inspection the few available resources prove to be less useful than they appear to be. Numerous online dictionaries which hold out the promise of providing useful information for low density language engineering are in fact short word lists. The few available electronic bilingual and multilingual dictionaries tend to be small, offering simple translations rather than word-meanings. Also, some applications which claim to contain dictionaries in numerous languages actually require the user to build the dictionary him/herself. In short, even the few electronic monolingual/multilingual dictionaries there are have notable drawbacks.

Provision of term-banks is equally poor. Term-banks are usually categorised according to a particular genre e.g. medical/legal/engineering and are extremely useful translation tools when working with specific types of texts. But the few available electronic term-banks for low density languages, like the available electronic dictionaries, is very small. For the few that exist, the range of genres covered is limited.

2.4. OCR Software

It is possible to scan text produced using any writing system as a graphics file (e.g. GIF or JPEG), and this is sometimes a method that web-publishers use to produce text to be displayed on a web-site. Yet optical character recognition software is necessary if scanned text is to be edited, or stored in a searchable corpus-based format. While OCR software rarely gives a 100% accurate rendition of a text, post-editing a piece of OCR data is a much quicker way of producing an electronic version of a printed text than typing it exclusively by hand. Hence access to OCR software is of importance to corpus construction for low density languages, especially as many of these languages have a low ambient level of electronic text available from publishers or on the web, increasing reliance on non-electronic texts in corpus building (Singh, McEnery and Baker, 2000). The lack of availability of OCR software, especially for Indic writing systems, is a major impediment to corpus construction for low density languages.

3. Reviewing the need for minority language engineering resources

Set against the context outlined in section two, the Engineering and Physical Sciences Research Council in the UK funded the Minority Language Engineering Project (MILLE⁵) to investigate the provision of language engineering tools and resources for non-indigenous minority languages (NIMLs) spoken in the UK. The project was undertaken by the Universities of Lancaster and Oxford, with the participation of a wide range of supporting academic, industrial and public sector partners.

⁵ Grant number GR/L96400.

The project focused on NIMLs, as indigenous minority languages use funding routes available to them via national governments, regional agencies and Europe to promote corpus construction. It was unclear, however, whether UK NIMLs such as Bengali, Hindi and Somali were being provided with language engineering resources. Such languages are important in the UK context, with a large volume of UK domestic translation being into such languages rather than, say, French or German. MILLE had the task of discovering which languages were being well provided with resources and which, of those spoken in the UK, were not. In addition, we wanted to develop a strategy for resource creation that fitted the needs of large numbers of language engineers in the UK and beyond working with low density languages. To this end, we decided to ask the language engineering community which languages they saw as being those for which language engineering resources were lacking, and which languages, given the necessary resources, they would like to work with.

In undertaking our review, we took the term language engineering community at its broadest level of meaning, incorporating those who are working both in the academic and commercial sectors.

3.1. The Questionnaire

Paper based questionnaires often receive poor response rates, possibly because of the administrative work that goes into their completion and return. Consequently, we decided to mount our questionnaire as a web-based html document. This saved on postage and printing costs, and allowed our respondents to access the document at their leisure. To further save time and effort for the respondents, we kept the questionnaire short (13 items) and mainly used check-boxes to save the respondents from having to type their replies. We also ticked a default answer of *no response* for each of the questions to reduce response times.

As we were aware that minority language engineering is still a relatively new area, we knew that if we asked only those people who were building or using minority language corpora to answer our questionnaire we would receive only a few responses. Therefore, we asked respondents to think about and anticipate future needs, and expanded our survey population to the language engineering community at large. In order to ascertain whether the demand for the resources we were asking people to consider was likely to arise, we included a question on the respondent's likelihood of working with low-density language corpora in the future.

We alerted the language engineering community to the existence of the questionnaire by sending messages to relevant language engineering/corpus linguistics email lists. We received sixty-seven responses to the questionnaire from research groups and individuals worldwide. Table one below shows the grouped nationalities of each respondent.

We asked each respondent which languages they would like to see corpus resources available for. As MILLE's focus was largely on UK NIMLs, we listed 13 UK NIMLs on a checklist (Arabic, Bengali, Chinese, Farsi, Gujarati, Hindi, Panjabi, Somali, Singhala, Sylheti, Tamil, Vietnamese and Urdu) though we left space for respondents to include languages from beyond this list.

Respondents were free to request as many or as few languages as they wished. The results are shown (in descending order of frequency) in table 2.

Location of respondent	Number of respondents
North America	20
Western Europe (excluding the UK)	11
UK	9
India	8
East Asia	6
Australia	4
Middle East	4
Eastern Europe	3
Africa	2

Table 1: Numbers and locations of respondents

Those for which the MILLE project found reasonable amounts of corpus data to be available, or under construction, are marked with an asterisk. What is notable about this table is that Indic languages are by far the most commonly requested resources. If the requests for Indic languages were collapsed into a combined score, they would clearly be the resources most frequently requested (with 82 requests).

Language	Number of requests for this language
Chinese*	28
Arabic*	19
Hindi	18
Vietnamese*	17
Tamil	15
Farsi*	11
Urdu	11
Gujarati	10
Bengali	9
Panjabi	9
Singhalese	6
Sylheti	4
Somali*	3

Table 2: Languages requested by the respondents.

3.2. Corpus Resources

The largest part of the questionnaire was concerned with the coverage and composition of the corpora requested. We asked what type of data the respondents wanted - monolingual or multilingual data (table 3).

Type of data requested	Number of requests
2 (bilingual)	19
1 (monolingual)	14
more than 2 (mutlilingual)	12
all of the above	12
any (not important)	8

Table 3: Types of data requested

For those who wanted multi- or bilingual corpora, we asked them which language(s) they would like the corpus to contain. Generally, people specified pairs of languages

which would be one NIML (e.g. Hindi - see above) plus one other. In 32 cases, English was the preferred choice of the other language, followed by German (3), Spanish (2) and French, Danish, Turkish, Hindi and Swedish (all 1 each). We also asked those who wanted multilingual or bilingual corpora to specify the level of alignment they would like between each language (if any):

Type of alignment requested	Number of requests
same texts -sentence aligned	24
same texts – word aligned	19
different texts - equivalent genres	8
same texts - no alignment	3
different texts - different genres	1

Table 4: Alignment requests

Regarding the content of the corpus, we asked whether transcribed spoken corpora or written corpora would be preferred (see table 5) and what the corpus balance should be (see table 6).

Written to spoken data ratio	Number of requests
both, but an emphasis on written	23
both	12
written only	6
both, but an emphasis on spoken	6
spoken only	1

Table 5: Written v. spoken data

Genre weighting requested	Number of Requests
balanced	34
either	17
focussed	13

Table 6: Genre balance

We then listed 12 genres common to available corpora and asked respondents to check which ones they would like to see featured in NIML corpora (table 7).

Genres requested	Number of requests
scientific	41
news	40
commerce	37
government	37
historical	34
fiction	33
manuals	33
legal	32
health	29
letters/diaries	29
leisure	26
children's	25

Table 7: Genres requested

We also allowed space for respondents to name other genres that we had not listed. The following answers were given: transcribed naturalistic conversations (5),

religion/spiritual (3), classics (2), narratives, botany, textbooks, non-native communication, cookery, poetry, financial, philosophy, banking, insurance, chemistry, websites, adverts and proverbs (1 each).

In summary, the corpora requested in the survey were very much like the British National Corpus but with a multilingual slant. Respondents were asking for large, balanced corpora, which provided more written than spoken data. There was an overwhelming demand for corpus data in a range of Indic languages. However, rather than simply wanting monolingual corpus data for these languages, there was clearly a wish that bilingual/multilingual corpus data be provided as well.

3.3. Corpus Encoding and Annotation

Having determined what types of corpus data in what languages were requested, we moved on to examine what encoding formats and annotations were requested by the respondents.

Annotation type	Number of requests
part-of-speech	43
parsed	31
phonemic	22
prosodic	16
semantic	36
no annotation (just plain text)	25

Table 8: Types of annotation requested

We offered a number of options for linguistic annotation (Garside, Leech and McEnery, 1998), and allowed respondents to check as many as they liked (see table 8). Other linguistic annotation that respondents requested related to morphology (3 requests), etymology, topic marking, pragmatics, error analysis, non-standard language use, code switching and marking theme/rheme (1 request each).

In addition we asked respondents to select their preferred character set encoding for NIML corpora.

Encoding format	Number of requests
Unicode	37
8-bit font based solution	25

Table 9: Character sets

With reference to the mark-up of the corpus texts, the findings of the Baker, Burnard, McEnery and Wilson (1998) survey of markup standards were supported, with a TEI based markup, encoding a minimal set of elements, being strongly preferred. However, a small number of respondents (6) requested CHILDES encoding.

With reference to transcribed speech corpora, the respondents indicated that, on balance, they would prefer transcription to occur in the native script of the speakers, with only 27 of the respondents requesting a romanised transcription for languages which do not use the Roman alphabet. The results for the preferred delivery format of the corpus is shown below.

Delivery format	Number of requests
World Wide Web	53

CD	39
ftp	32
diskette	18
dat tape	5

Table 11: Delivery formats requested

3.4. Proposed Applications

We asked the respondents to imagine that they had a CD of corpus data for a range of European NIMLS of both written and spoken language. We then asked them what sort of applications they would have for such a corpus. The most common answers were: dictionary and vocabulary list construction (12), machine translation (9), developing semantic annotation tools (7), exploring issues related to code-switching (6), developing teaching aids (6), developing computational grammars (6), exploring differences in genre/context (4), exploring issues relating to phonology (4), developing speech recognition systems (3), building models of discourse (2), exploring prosody (2), producing text-to-speech tools (2), developing spell-checkers (3) and information retrieval (2).

We asked everybody to name the kind of support tools they would need in order to exploit this imaginary corpus data. The most frequently listed tools were concordancers (13) mark-up tools (4), frequency lists/counts (4), dictionaries and dictionary builders (3), alignment tools (2), text-editors (2), translation-based tools (2).

Finally, we asked each respondent how likely they were to be working with NIMLS in the future. The results are shown in the table below:

Probability of working in minority language engineering in the future	Number of responses
Very likely	41
Possibly	10
Probably not	9
Unsure	5
Very unlikely	1

Table 12: The likelihood of respondents working in minority language engineering

The answers from the language engineering questionnaire enabled us to build a portrait of an idealised NIML corpus, based on the demands of the language engineering community. It was clear that a collection of balanced monolingual spoken and written Indic language corpora would be attractive to potential users, especially if parallel corpora for the languages was also made available. In terms of encoding, the Baker, Burnard, McEnery and Wilson (1998) recommendations should be followed, and the corpus should use Unicode. If any linguistic annotation is to be introduced it should at least cover part-of-speech information, and sentence alignment of the parallel texts is desirable.

All indications are that, if such corpora were built there are potential users waiting to use them to generate a wide range of language engineering applications.

Based upon this review, a new project has been initiated to provide just such corpus data.

4. Enabling Minority Language Engineering

On the basis of the findings of MILLE, a new project Enabling Minority Language Engineering (EMILLE) has been funded by the EPSRC in the UK⁶. EMILLE is designed to address a range of issues to enable language engineering research on Indic languages. The project will construct 9,000,000 word written corpora (including both monolingual and parallel data) and 500,000 word spoken corpora for Bengali, Gujarati, Hindi, Panjabi and Urdu. These are the major UK Indic NIMLS (see Baker *et al.*, 1999, for a description of UK NIML communities). As the review of the needs of language engineers world-wide presented here also found a need for Tamil and Singhalese corpora in the language engineering community, we have also undertaken to produce 9,000,000 word written corpora for these languages. However, as neither are major UK NIMLS, we will be unable to gather spoken corpora for these languages.

To enable the work which language engineers have indicated that they would like to do with such corpora, the project will also focus on establishing a language engineering architecture within which minority language engineering may take place. Language engineering architectures need to expand beyond their current focus on European languages. To be truly generic platforms, language engineering architectures cannot be limited to specific languages/writing systems. To this end, EMILLE will extend GATE to be fully UNICODE compliant so that it may act as a framework within which the corpora of EMILLE can both be developed and exploited⁷. Within the GATE framework tools will be developed to allow for mapping a diverse range of font-based representations of Indic writing systems into UNICODE. The project will also undertake the part of speech tagging of at least one of the languages represented in the corpus in both spoken and written form. Finally, the project will develop existing alignment software to sentence align the parallel corpora within EMILLE. This alignment facility will be embedded within the GATE architecture.

Our strategy is to work on major languages which a large number of researchers wish to work with but currently cannot. Yet in focusing on developing a language independent language engineering architecture as well as corpus data for Indic languages, we hope to help as many researchers working on low-density languages as possible. By providing generic solutions to problems faced by Indic languages, we hope to enable work in a wider range of languages.

5. References

- Baker, J.P., Burnard, L., McEnery, A.M. and Wilson, A. "Techniques for the Evaluation of Language Corpora: a report from the front", *Proceedings of the First International Conference on Language Resources and Evaluation*, Spain, pp 135-142.
- Baker, P. and McEnery, A. *Needs of language-engineering communities; corpus building and*

⁶ Grant number GR/N19106.

⁷ In doing so, we will be casting a wider net than other related efforts. For example, Pangea at New Mexico Computing laboratory has goals similar to that of EMILLE – but circumvents Indic languages. See <http://crl.nmsu.edu/Research/Projects/pangea/index.html> for details.

- translation resources*. MILLE working paper 7, Lancaster University, 1999.
- Baker, P., Burnard, L., McEnery, A. and Sebba, M. *Locating minority language-speaking communities in the UK*. MILLE working paper 3, Lancaster University, 1999.
- Chareoeonporn, T. (ed) *Technical report: ORCHID Corpus*, National Electronics and Computer Technology Centre, Thailand, 1997.
- Garside, R., Leech, G. and McEnery, A. *Corpus Annotation*, Longman, London, 1998.
- Hall, P.A.V. and Hudson, R. (editors) , *Software without Frontiers*, John Wiley and Sons, New York, 1997.
- Lie, M., Baker, P., McEnery, A. and Sebba, M. "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, 1999.
- Nirenburg, S. and Raskin, V. "Universal Grammar and Lexis for Quick Ramp-Up of MT Systems". *Proceedings of ACL/COLING '98*. Montréal: University of Montreal, 1998.
- Ostler, N. "Language technology and the Smaller Language", *ELRA Newsletter*, 4 (2), 1999.
- Singh, S., McEnery, A. and Baker, P. "Building a parallel corpus of Panjabi-English", in J. Veronis (ed), *Parallel Text Processing*, Kluwer, Dordrecht, 2000.
- Somers, H. "Machine Translation and Minority Languages", *Translating and the Computer 19: Papers from the Aslib conference*, London, 1997.