Corpus resources and minority language engineering

Tony McEnery, Paul Baker & Lou Burnard

Department of Linguistics, Lancaster University
Bailrigg, Lancaster, LA1 4YT
mcenery@comp.lancs.ac.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 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. Hence the lack of corpus data for a language may have severe consequences for the future of that language, for as Ostler (1999:3) states "Languages which do not take a full part in the electronic media are doomed to stagnate, if not atrophy". This is a state of affairs which has long been recognized at Lancaster University, and is one which we have responded to. From our early work in English corpus linguistics, we have moved on to examine European language, such as French, Spanish (REF) and Polish (REF), and Far Eastern languages such as Chinese (REF). Recently, we have continued this work by focusing on South Asian languages, with work on Panjabi (REF) and Sylheti (REF). This paper outlines how we have developed our strategy for working on South Asian langauges. In this paper, we would like to argue that, based upon our work to date, it is with these languages that a real need for corpus data exists in the language engineering community that is simply not being met at this moment in time. This paper is about why this need should be met, and how we intend to meet it. As such the paper has three main goals. Firstly, we will review the state of language processing technology for low density languages, updating the work of Somers (1997). Secondly, we want to present the findings of a major review (with over 80 research centres world-wide participating) of the needs of language engineers in relation to low density languages. Finally we will present a summary of the technical problems faced by those developing LRs for low density languages and propose solutions to them. In doing so, we will present proposals to extend a current language engineering architecture, GATE, to act as an architecture for low-density language engineering.

2. The State of Minority Language Engineering

Somers (1998:6) gives a table which lists the availability of various resources for different "exotic" languages (adapted from Hearn 1996 and World Language Resources 1997). 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. Spell-checkers, for example, were only available for Arabic.

The computational landscape changes rapidly, however, and an updated version of the table, carried out in 1999 shows a less gloomy state of affairs. The situation is still relatively bleak, however for Indian languages, especially Sylheti, which in its written form can reasonably be viewed as an endangered language (REF).

2.1. Word processing, hyphenation and fonts

It is possible to find fonts in almost all languages now, even African and Indian languages. A good-quality font, used with a word processor such as Microsoft Word is generally the solution that most of the translators we questioned now employ. Working with fonts can be problematic, notably because a key-mapping must be learnt, and different font-sets of the same language can use different mapping systems. Also, provision for diacritics and conjunct characters in Indian-based languages can be patchy.

A few language-specific or multilingual word-processors are available i.e. those which include menus and help in multiple languages. A number of companies now sell add-on multilingual hyphenation software e.g. Hyphenation for Ventura, Hyphenologist, although multilingual word-processors are becoming increasingly sophisticated and are incorporating hyphenation rules as part of the package.

2.2. Dictionaries and term banks

The table only includes electronic provision of dictionaries, whether available via the internet or by computer software. The phrase "dictionary" is misleading as there are numerous examples of online dictionaries which are in fact short word lists. Somers (1997:6) notes that dictionaries are usually more than word-lists; they offer some grammatical information too. At what point does a wordlist become "useful"? A list of 100 words might not be of much use for translators. Bilingual and multilingual dictionaries tend to be small, offering simple translations rather than word-meanings. Also, provision of bilingual and multilingual dictionaries in the table is restricted to specific languages. A bilingual French-German dictionary will be of no use to someone who wants to translate French to English. Also, some applications which claim to contain dictionaries in numerous langugaes actually require the user to build the dictionary him- or her-self.

Provision of term-banks can be equally patchy. 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. Again, the provision of a particular term-bank for a language in the table does not imply that all genres are represented, and it also should be noted that electronic term-banks, like dictionaries can be very small.

2.3. OCR Software

It is possible to scan text from any language as a graphic, and this is sometimes the method that web-publishers use to display text in foreign languages. However, optical character recognition software is neccesary if the text is to be edited, or stored in a searchable corpus-based format. Although most romanised scripts are now dealt with by OCR software, and Chinese also has some provision, there is still a gap in the market for Indian languages which use Devanagari and Gurmukhi scripts. OCR software rarely gives a 100% accurate rendition of a text, but post-editing a piece of OCR data is much quicker than typing it by hand. Hence, the provision of OCR software for Indian languages would greatly enhance the feasibility of creating Indian corpora.

2.4. Unicode

Unicode, the 16-bit character set is perceived by many language engineers as the future for multilingual encoding. It is envisaged that all electronic text will eventually be formatted in Unicode or a format similar to it, alleviating the need for fonts and mapping tools. Hence, it is important that the character set of a minority language is fully represented in Unicode, as 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 version 3.0 is still in its developmental stages, although it is expected that a number of significant additions will appear on its release (Cherokee, Burmese, Canadian syllabics, Ethiopic, Maldivian, Sinhala, Khmer and Yi). As yet there is no provision for the Sylheti script Nagri.

3. Reviewing the need for minority language engineering resources

We take the term language engineering community at its broadest level of meaning, incorporating those who are working both in the academic and commerical sectors; for the purposes of our questionnaire we were happy to include anybody who uses or builds corpus-based resources in order to study language. A previous project carried out by members of MILLE had focussed on establishing guidelines for corpus annotation schemes (Baker, Burnard, McEnery & Wilson 1998), part of which had involved a survey of 26 corpus users. It was decided to build upon this earlier work in order to ascertain the needs of the language engineering community, but to focus on issues surrounding corpus building of European minority languages, rather than corpus building *per se*.

3.1. The Questionnaire

Based on previous experience of questionnaire design and implementation, it was decided to mount the questionnaire as a web-based html document. This would save on postage and printing costs, and allow our users to access the document immediately, and email their replies at the click of a button. Questionnaires often receive poor response rates, possibly because of the administrative work that goes into their completion and return - we considered the 50% response rate from our translator's questionnaire to be good! We also predicted that our intended respondents *would* be likely to have internet access, and have experience at filling in forms on the web. A 13 item questionnaire was mounted on a website, mainly using check-boxes to save the respondents from having to write their replies. For a large section of the questionnaire we ticked a "default" answer of "no response" for each of the 34 parts of this question, again to reduce reponse times.

As we were aware that minority language engineering is still a relatively new discipline of computational linguistics, we knew that if we asked only those people who were building or using minority language corpora to answer our questionnaire that we would only receive a few responses. Therefore, we asked members of the language engineering community to imagine that they would be building or using such corpora in the future and to answer the questions with this in mind. We also included a question on the respondent's likelihood of working with minority corpora in the future. We alerted the language engineering community to the existence of the questionnaire by sending messages to a number of specialist email mailing lists which dealt in linguistics, corpora and encoding, and as an extra incentive, we offered all respondents a free copy of our findings.

We received sixty-seven email responses to the questionnaire, more than twice as many as the translator's questionnaire or the ELRA survey. The table below shows the grouped nationalities of each respondent.

Location of Respondent	Number from Location
North America	19
West Europe (not UK)	9
UK	8
India	7
East Asia	5
Australia	3
Turkey	2
Eastern Europe	1
Africa	1
USSR	1
Iran	1

Table 1: Numbers and locations of respondents

We asked each respondent which language they would like to see corpus resources available for. We listed 13 (mainly) UK NIMLs (Arabic, Bengali, Chinese, Farsi, Gujarati, Hindi, Panjabi, Somali, Sinhala, Sylheti, Tamil, Vietnamese and Urdu) but also left space for respondents to choose their own language. The results are shown (in order of preference) in the table below:

Language	**n**
Chinese	28
Arabic	19

Hindi	18
Vietnamese	17
Tamil	15
Farsi	11
Urdu	11
Gujarati	10
Bengali	9
Punjabi	9
Sinhalese	6
Sylheti	4
Somali	3

Table 2: Languages for which a need for language engineering resources was identified.

3.2. Corpus Resources

The largest part of the questionnaire was concerned with the encoding of NIML corpora.

Number of languages in corpus:

number of languages	**n**
1 (monolingual)	14
2 (bilingual)	19
more than 2 (mutlilingual)	12
any (not important)	8
all of the above	12

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 mult- or bilingual corpora to specify the level of alignment they would like between each language (if any):

alignment	
same texts - word aligned	19
same texts -sentence aligned	24
same texts - no alignment	3
different texts - equivalent genres	8
different texts - different genres	1

Regarding the content of the corpus, we asked whether spoken (transcribed) or written corpora would be preferred:

written:spoken ratio	
written	6
spoken	1
both	12
both, but an emphasis on written	23
both, but an emphasis on spoken	6

We also asked whether the corpus should be balanced across genres, or focus upon one genre:

Genre weighting	
balanced	34
focussed	13
either	17

We then listed 12 genres (health, legal, news, government, leisure, commerce, scientific, fiction, children's, historical, letters/diaries, manuals) and asked respondents to check which ones they would like to see featured in NIML corpora:

genre	**number**
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

We had also allowed space for respondents to name other genres that we had not explicity listed. The following answers were given: transcribed naturalistic conversations (5), religion/spiritual (3), classics (2), narratives, botany, textbooks, non-native communication, cookery, poetry, financial, philosophical, banking, insurance, chemistry, "period", websites, adverts, proverbs, literary (1 each).

3.3. Corpus Encoding and Annotation

We offered a number of options for linguistic annontation, taken from Corpus Linguistics (McEnery & Wilson 1996), and allowed respondents to check as many as they liked:

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

Other types of annotation that respondents would have liked to see employed in NIML corpora were morphological (3), etemology, topical, pragmatic,

linguistic errors, non-standard language, mixed langugaes and theme/rheme (1 each).

We also asked respondent to name their preferred encoding format(s) for NIML corpora.

Encoding format	**number**
Unicode	37
8-bit font	25
romanised transliteration scheme	27

Preferred delivery format of the corpus is shown below:

Delivery format	**number**
diskette	18
cd	39
ftp	32
dat tape	5
World Wide Web	53

We then asked which form(s) of textual markup would be preferable:

Textual annotation	number
TEI-Lite	7
TEI	8
SGML	13
XML	23
HTML	24
CHILDES/LIDES	6

We next presented the respondents with a list of encoding features: header elements, primary data, paragraph level elements, and spoken data, and asked them to mark the importance of encoding each feature in a NIML corpus. We had specified the default option to be "no opinion". Each answer was awarded 1 if marked "essential", 2 if marked "if possible", 3 for "no opinion", and 4 for "not wanted", enabling us to calculate a mean result for each feature where 1 would be most preferable and 4 would be least preferable. The results are presented in order of preference in the table below:

3.4. Proposed Applications

We told linguists and language engineers to imagine that they had a cd of corpus data for a range of European non-indigenous minority languages in both written and spoken formats. We then asked them what sort of questions they would want to explore with such a corpus.

The most common answers for linguists were questions involving: machine translation (9), dictionary and vocabulary building (7), teaching aids (6), speech recognition (3), text-to-speech (2), spell-checkers (3), computational grammars (2), information retrieval (2).

For language engineers, the most common answers involved: semantics (7), language contact issues (loanwords, dialect, code-switching) (6), syntax (4), differences in genre/contexts (4), phonology (4), frequencies (3), dictionary and vocabulary building (2), interpersonal/discourse (2), prosody (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 (8), search tools (5), 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:

probablity	n
very likely	41
possibly	10
unsure	5
probably not	9
very unlikely	1

The answers from the language engineering questionnaire enables us to build a portrait of an idealised NIML corpus, based on the most popular answers. Such a corpus would most likely be a language pair such as Chinese-English, Arabic-English or Hindi-English. It would be aligned at the sentence level, consisting mainly of written texts, with a smaller spoken section. It would be balanced across several genres - most likely containing texts from science, news, commerce and government domains, and would be part-of-speech annotated. The encoding format would be Unicode, and it would be available over the World Wide Web, marked up using html or xml. Some header information would be contained, while annotation of the text would at least be at the paragraph level.

4. Enabling Minority Language Engineering

EMILLE is designed to address a range of issues to enable language engineering research on Indic[1] 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, see Reynolds, 1996, for evidence of the permanence of these languages in NIML communities in the UK). As the review of Baker & McEnery (1999) also found a need for Tamil and Singhalese corpora in the language engineering community, we will undertake to produce 9,000,000 word written corpora for these languages also. However neither are major UK NIMLs, we will be unable to gather spoken corpora for these languages.

The project will also focus on establishing a language engineering architecture within which minority language engineering may take place. The EPSRC workshop on

[1] 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. EMILLE, however, is concerned with a sub-set of Dravidian and Indo-Aryan languages only.

language engineering architectures[2] led to a discussion focused around the need of language engineering architectures 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 exploited[3]. 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.

5. References

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 252--262). Hillsdale, NJ: Lawrence Erlbaum Associates.

[2] "A Workshop on Language Processing Architectures and the Use and Distribution of Language Resources", EPSRC ref. GR/M44545.

[3] 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.